AL-TP-1992-0006

ARMSTRONG LABORATORY

TECHNIQUES TO INFER JOB COMPETENCY LEVELS FROM HANDS-ON JOB PERFORMANCE SCORES

Carolyn H. Fotouhi
Gregory P. Mosher
Dickie A. Harris

Human Resources Research Organization
1100 South Washington Street
Alexandria, VA 22314

Donald L. Harville
Mark S. Teachout

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

March 1992

Final Technical Paper for Period March 1990 - May 1991

Approved for public release; distribution is unlimited.

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

NOTICES

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

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

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

This paper has been reviewed and is approved for publication.

DONALD L. HARVILLE
Contract Monitor

HENDRICK W. RUCK, Technical Director
Technical Training Research Division

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
March 1992

3. REPORT TYPE AND DATES COVERED
Final March 1990 – May 1991

4. TITLE AND SUBTITLE

Techniques to Infer Job Competency Levels from Hands-on Job Performance Scores

5. FUNDING NUMBERS

PE - 63227F
PR - 2922
TA - 01
WU - 05

6. AUTHOR(S)

Carolyn H. Fotouhi
Gregory P. Mosher
Dickie A. Harris
Donald L. Harville
Mark S. Teachout

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

Human Resources Research Organization
1100 South Washington Street
Alexandria, VA 22314

8. PERFORMING ORGANIZATION REPORT NUMBER

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

Armstrong Laboratory
Human Resources Directorate
Technical Training Research Division
Brooks Air Force Base, TX 78235-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

AL-TP-1992-0006

11. SUPPLEMENTARY NOTES

Armstrong Laboratory Technical Monitor: Donald L. Harville, (512) 536-2932.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The background of the Job Performance Measurement (JPM) project and the performance measures collected by that project are described. Job performance test scores from the Job Performance Measurement System need to be interpreted in terms of job competence. The literatures on item-based and examinee-based standard setting methods are reviewed and compared. Subject matter experts use a modified Jaeger technique to set standards for 16 hands-on tasks for the Aerospace Ground Equipment AFS. A technique using archival JPM scores and test administrator ratings of task performance also is used to set standards. Both methods result in considerable savings of time and money compared to traditional standard setting methods. The standards resulting from using those two techniques are similar. Implications for future standard setting research and practice are discussed.

14. SUBJECT TERMS

Job Competency
Performance measures
Personnel
Personnel selection
Standard setting
Work samples

15. NUMBER OF PAGES
68

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
UL

TABLE OF CONTENTS

Introduction 1

JPM Performance Measures 2

Standard Setting Methods 5

Item-Based Methods 5
Examinee-Based Methods 9
Comparisons of Item-based and Examinee-based Methods 11
Standard Setting Process 13

Setting Job Performance Standards for the JPM Hands-on Tests 16

Item-Based Judgmental Technique 17

Method 18

Participants 18
Procedure 18
Procedure 22
Conclusions 24

Examinee-Based Archival Technique 29

Method for Calculating Cutoffs 31
Results and Discussion 33
Conclusions 48

Comparison of Archival and Judgmental Standards 49

Topics for Further Consideration 52

Selection and Classification 52
Training Evaluation 53

References 55

TABLE OF CONTENTS (continued)

List of Tables

1. Hands-on Test Code Numbers and Test Titles 4
2. Tests Discussed at Each Base 23
3. Judgmental Technique: Minimum Test Score Cutoff by Test and Round . 25
4. Planned Comparisons for Round Effects on Cutoff Means and Standard Deviations Using Repeated Measures ANOVA 28
5. Reliability of Hands-on Tests by Round 29
6. Methods of Categorizing Examinees into Five Proficiency Levels and Calculating Mean Hands-on Test Score (MHOTS) at Each Level 32
7. Example of Adjustment for Missing Data for Test 154--Perform an Aircraft Support Generator Service Inspection 33
8. Example of Adjustment for Missing Data for Test 179--Perform a Gas Turbine Compressor Periodic Inspection 34
9. Archival Technique: Minimum Test Score Cutoff by Test and Classification Method 35
10. Planned Comparisons for Method Classification Using Repeated Measures ANOVA 36
11. Archival Technique: Correlation of Test Administrator (TA), Supervisor (Super), Incumbent (Incum), and Peer Ratings with Hands-on Test Score 38
12. Results of Multiple Factor ANOVAs of Hands-on Test Score by Rater . 40
13. Results of the Linear Regression of Mean Hands-on Test Score and the Individual Raters 45
14. Archival and Judgmental Techniques: Minimum Test Score Cutoff by Test 50

List of Figures

1. Proficiency levels for setting job performance standards 17
2. Standard setting form for Test 154--Perform an Aircraft Support Generator Service Inspection 20
3. Standard setting form for Test 154--Perform an Aircraft Support Generator Service Inspection 21
4. Judgmental technique: Standards across all hands-on tests for Round 1 and Round 2 27
5. Calculation of mean hands-on test scores (MHOTSs) by rater and proficiency level 31
6. Archival technique: Standards across all hands-on tests set via three classification procedures 36
7. Archival technique: Technical Proficiency (TP) level by hands-on test score via separate sources for Test 154--Perform an Aircraft Support Generator Service Inspection 37
8. Archival and judgmental techniques: Standards across all hands-on tests 51

PREFACE

The research described in this paper was funded through MIPR #FQ7624-90-00024 M0010 to the Defense Training and Performance Data Center (TPDC) which was used for contract M67004-89-D-0001 with Star Mountain, Inc. The Human Resources Research Organization (HumRRO) was the sub-contractor. The purpose of the research described in this paper was to interpret job performance test scores from the Job Performace Measurement System (JPMS) in terms of job competence.

SUMMARY

The paper describes the background of the Job Performance Measurement (JPM) project and the performance measures collected by that project. The literatures on item-based and examinee-based standard setting methods are reviewed and compared. Subject matter experts use a modified Jaeger technique to set standards for 16 hands-on tasks for the Aerospace Ground Equipment (AGE) AFS. A technique using archival JPM scores and test administrator ratings of task performance also is used to set standards. Both methods result in considerable savings of time and money compared to traditional standard setting methods. The standards resulting from using those two techniques are similar. Implications for future standard setting research and practice are discussed.

TECHNIQUES TO INFER JOB COMPETENCY LEVELS FROM HANDS-ON JOB PERFORMANCE SCORES

Introduction

The Joint-Service Job Performance Measurement/Enlistment Standards (JPM) Project was initiated by a Congressional mandate in 1980. In response to the mandate, each branch of the Services launched separate, but related, research projects. The primary goal of the JPM Project was to develop better methods for assessing the job performance of enlisted personnel. If improved assessment methods could be developed, a secondary goal of the project was to explore the feasibility of linking job performance to enlistment standards. To accomplish the first goal, the Air Force Human Resources Laboratory (AFHRL) developed and administered different types of performance measures to first-term airmen in various occupations, or Air Force Specialties (AFSs). The performance measures consist of hands-on tests; interview tests; rating scales; and written, multiple-choice job knowledge tests. These measures were developed for and administered to first-term airmen in eight AFSs: (a) Air Traffic Control, (b) Avionic Communications, (c) Aircrew Life Support, (d) Precision Measuring Equipment Laboratory (PMEL), (e) Information Systems Radio, (f) Jet Engine Mechanic, (g) Personnel Specialist, and (h) Aerospace Ground Equipment (AGE). The results from this research effort indicated that the newly developed performance measures were an improvement over existing assessment methods, which only partially assessed job performance (Office of the Assistant Secretary of Defense, 1989).

The primary goal of the JPM Project, the development of improved job performance measures, has been realized. The next step is to explore the feasibility of establishing a link between job performance and enlistment standards. However, before this second goal can be achieved, job performance standards must be established.

Wigdor and Green (1986) stated in the evaluation of the JPM Project that "to be really useful in the central matter of setting standards and allocating recruits among job specialties, the project's primary measurement goal should be to supply performance scores with some absolute meaning, i.e., to measure individuals' proficiency with reference to the whole job. This we have designated as a competence approach" (pp. 55-56). Attaching meaning to performance scores will give policy makers a better sense of an incumbent's total job capability and of the practical benefits associated with higher selection and classification standards in terms of performance and costs. To interpret performance test scores in terms of job competence, the representativeness of the test must be established. Once this is determined, the test can be scored to represent the amount of, or the extent to which the job content domain has been mastered (Wigdor & Green). This requires that absolute meaning be given to individual performance scores or to the distribution of performance scores, so that test scores can be interpreted in terms of what an incumbent knows or can do, rather than how the incumbent compares with other incumbents. By ascribing meaning to the performance measures (e.g., hands-on performance tests [HOPT]) and the resulting performance scores and by referencing them to an external scale of job requirements, job competency or proficiency can be more readily determined.

Wood and Power (1987) draw an interesting distinction between competence and performance: "Competence refers to what a person knows and can do under ideal circumstances, whereas performance refers to what is actually done under existing circumstances. . . . Developed competence is to be conceived of and assessed as a continuous variable reflecting various degrees of integration of knowledge and skill, of understanding and proficiency" (p. 415).

In line with the stated goals of the JPM Project is the view that, to derive the most useful information from the performance measures, an absolute meaning must be attached to a performance score. Such an absolute meaning can be established with the measurement of competence or job proficiency. As Glaser (1963) wrote:

> Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance. An individual's achievement level falls at some point on the continuum as indicated by the behaviors he displays during testing. The degree to which his achievement resembles desired performance at any specified level is assessed by criterion-referenced measures of achievement or proficiency. The standard against which a student's performance is compared when measured in this manner is the behavior which defines each point along the achievement continuum. The term "criterion," when used in this way, does not necessarily refer to final end-of-course behavior. Criterion levels can be established at any point in instruction as to the adequacy of an individual's performance. The point is that the specific behaviors implied at each level of proficiency can be identified and used to describe the specific tasks a student must be capable of performing before he achieves one of these knowledge levels. It is in this sense that measures of proficiency can be criterion-referenced.
>
> Along such a continuum of attainment, a student's score on a criterion-referenced measure provides explicit information as to what the individual can or cannot do. Criterion-referenced measures indicate the content of the behavioral repertory, and the correspondence between what an individual does and the underlying continuum of achievement. Measures which assess student achievement in terms of a criterion standard thus provide information as to the degree of competence attained by a particular student which is independent of reference to the performance of others (p. 519).

JPM Performance Measures

The goal of the current paper is to establish minimum job performance standards for one of the eight AFSs under study in the JPM Project--the Aerospace Ground Equipment Mechanic (AGE). If the procedure results in acceptable standards for one AFS, it can be used to set standards for other AFSs. The AGE AFS was selected for initial study because the JPM data were collected more recently and on a greater number of airmen than for the other seven AFSs in the JPM Project.

There are four instruments available for establishing job performance standards: (a) job knowledge tests, (b) hands-on tests, (c) interview tests, and (d) rating scales. The rating scales are comprised of continuously scored items; the remaining measures are scored dichotomously. The job knowledge tests follow a written, multiple choice format, whereas the hands-on tests are performance or worksample tests scored YES/NO. The interview tests are a type of performance test and were developed to assess tasks that are important to the AFS but are too expensive, too time-consuming, too dangerous, etc. to assess with a hands-on measure. The examinee talks through the procedures necessary to perform the particular task and may be prompted by the interviewer/scorer. The interview tests, like the hands-on tests, are scored YES/NO. Of the four job performance measures developed for the JPM Project, the hands-on tests have the most face validity (Office of the Assistant Secretary of Defense, 1989). Therefore, it was decided to set standards on the hands-on tests. If needed, standards can be set for the other job performance measures at a later time.

Because the job performance measures were developed as part of a selection and classification study, it is reasonable to assume that they adequately cover the AFS under consideration (Lipscomb & Dickinson, 1987). A stratified random sampling procedure was used for task selection and sampling. The goal was to define the job domain of interest--tasks performed by first-term airmen. Occupational survey information was analyzed to define the domain of interest based on rational stratification factors, such as the proportion of first-termers who performed the task and the relative percent of time they spent on those tasks. In the last step of the process, subject matter experts (SMEs) eliminated tasks that were no longer performed, required equipment such as an airplane that would be impractical to dedicate to testing, required identical or almost identical abilities to perform as a task already included, or were not observable or measurable. In the cases where a task was eliminated by the SMEs, another task was selected randomly from the pool of tasks identified from the stratified random sampling procedures.

Sixteen tasks, which are common across the specialty, were selected for hands-on testing in the AGE AFS. A list of the tests and their corresponding code numbers is presented in Table 1. The tests range in length from 7 to 30 items or steps, and each step is scored YES/NO in terms of whether the examinee correctly performed the step. Examinees are required to perform the hands-on test according to technical order (T.O.) specifications and are permitted to reference T.O. manuals, workcards, and other written materials that are regularly used in performing the task. The tests are administered in the workplace by an AFHRL trained test administrator who is unknown to the examinee. The test administrator observes the examinee perform the test and scores each step YES/NO as it is/is not performed.

A weighting system is used in computing hands-on test scores in which steps that are more difficult or more critical to successful completion of the test receive more weight than easier or less critical steps. Criticality weights were assigned by a group of senior Non-Commissioned Officers (NCOs)

Table 1

Hands-on Test Code Numbers and Test Titles

Test Code	Test Title
154	Perform an Aircraft Support Generator Service Inspection
155	Perform a Service Inspection on a Load Bank
162	Perform a Service Inspection on a Hydraulic Test Stand
179	Perform a Gas Turbine Compressor Periodic I.ispection
209	Measure Resistance in AGE Electrical Systems
215	Perform AGE Electrical Systems Operational Checks
238	Splice Electrical Systems Wiring
251	Adjust Turbine Engine Fuel System Components
260	Clean Motor and Generator Components
264	Isolate Engine, Motor, or Generator Malfunctions
284	Remove and Replace Engine Fan Belts
300	Remove or Install Fuel Line and Fittings
421	Remove or Install Hydraulic Lines
446	Isolate Pneumatic System Malfunctions
503	Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance Information
549	Inspect Vehicles for Safety of Operation

during scoring workshops held prior to data collection. Criticality weights are summed across all steps of a test to define the "base score" for that particular test. To calculate an individual's hands-on test score, weights for each step scored YES are summed, divided by the base score for that test, and multiplied by 10. Using this system, hands-on test scores for each task range from 0.00 to 10.00. By equating test scores on a 10-point scale, comparisons can be made across tests or a composite hands-on test score can be computed by summing or averaging the individual test scores.

Given that the hands-on tests adequately cover the AGE AFS (Lipscomb & Dickinson, 1987), the next question is whether to set standards on each test (e.g., each hands-on test) or to set standards on dimensions of job performance (e.g., all tests covering mechanics). Regardless of the level at which standards are set (i.e., test vs. dimension), these standards must be aggregated to form a standard for job performance. The methodology for aggregating standards is beyond the scope of the present project. Standard setting procedures are basically the same whether standards are set at the test or dimension level; therefore, the following discussion is limited to the identification of an appropriate procedure for establishing standards at the test level. Because initial efforts will focus on establishing performance standards for the hands-on tests, the ensuing discussion will highlight standard setting issues and concerns as related to those tests.

Standard Setting Methods

Standard setting procedures can be divided into two categories: (a) item-based and (b) examinee-based. Item-based methods require that raters make judgments regarding the proportion of minimally competent individuals who would correctly answer each test item. Proportions are aggregated across items and judges to form a "percent correct" standard. Examinee-based methods require that raters identify competent and noncompetent or borderline competent individuals who are then administered the test. Standards are then set based on the data obtained from the test administration. As can be seen from these general descriptions, all standard setting procedures require some subjective judgments. With item-based approaches, judges are required to make decisions about the test items. With examinee-based approaches, decisions are required concerning individuals who will be administered the test.

Item-Based Methods

In setting performance standards, item-based methods are more widely used than examinee-based methods. Perhaps because they are so widely used, the item-based methods have become synonymous with the names of their respective developers. The four item-based standard setting methods discussed in the following section are (a) Nedelsky, (b) Angoff, (c) Ebel, and (d) Jaeger.

Nedelsky method. The Nedelsky method requires a multiple choice test format. For each item, judges identify the distractors that a minimally competent individual would readily eliminate as incorrect. A minimum passing level (MPL) is then calculated for each item. The MPL is equal to the reciprocal of the remaining response options, after eliminating easily identifiable incorrect options. For example, if a judge identifies two of five distractors as easily eliminated by a minimally competent individual, the MPL is 1 divided by 3 or .33. Thus, the MPL is calculated for each item for each judge. A cutoff score is obtained for each judge by summing the MPLs across items for that judge. A standard for the test is obtained by averaging MPLs across judges.

To avoid classifying as incompetent an examinee whose true performance is just equal to the test standard, solely as a result of measurement error, Nedelsky (as cited in Jaeger & Keller-McNulty, 1986) recommends a downward adjustment of the initial standard. Working from several assumptions, the adjustment requires reducing the initial standard by one or more standard deviations of the distribution of MPLs obtained from the sample of judges.

The Nedelsky method is the most widely used method for setting standards for professional certification and licensure exams (Livingston & Kastrinos, 1982). However, there are several disadvantages to the method. The first disadvantage lies in the assumptions regarding examinee decision making processes. Once obviously incorrect options are identified, examinees are assumed to have no information, however partial it may be, on which to select from among the remaining response options. Therefore, it is assumed that examinees randomly choose among response options they cannot clearly identify as incorrect. In reality, single test items are not presented in a vacuum as these assumptions lead one to believe. Information from one item may be, and often is, used to help answer another item. The Nedelsky method does not account for this. Thus, resulting standards may be more lenient than intended.

Poggio (1984) summarizes several shortfalls with the method based on successive implementations with the Kansas minimum competency testing program. He found that raters were often confused by the method, and as a result, they reported not being confident in their judgments. Raters also tended to be careless in studying items and often designated the correct response as a viable distractor. Because the method is confusing, highly trained raters are required. While it is imperative that raters be experts in the area in which standards are to be set, it seems wasteful to spend extra time and resources training them on how to use a particular standard setting method when another method will work effectively without such extra training.

A primary disadvantage of the Nedelsky method given the goals of the JPM Project is that it can be used to set standards only on multiple choice tests. Therefore, it could be used to establish standards on the written job knowledge tests but not for any other job performance measures (i.e., hands-on tests, interview tests, and rating scales).

Angoff method. The Angoff method asks raters to think of a group of minimally competent individuals rather than only one person. Raters estimate the percentage of minimally competent individuals who would be able to answer each item correctly. The cutoff score for a particular rater is the sum of his or her percentages across items. The test standard is the average of cutoff scores across raters. Thus, the percentage of minimally competent individuals passing an item is converted to the percentage of items that should be passed by minimally competent individuals.

Compared to other standard setting procedures, the Angoff method is the most straightforward and the easiest to implement. Raters have essentially no problem understanding the task they are to perform.

One disadvantage is the amount of variability in the standards provided by the Angoff method (Poggio, Glassnap, & Eros, 1981). Variability is particularly a problem when only a few raters are used as is the case in most workshop settings. Jaeger and Busch (1984) used an iterative approach and normative data in an attempt to reduce the variability in ratings obtained via the Angoff method. Raters first provided independent ratings. After the presentation of normative data and a discussion period, raters were allowed to independently reconsider their original ratings. While the mean standard did not change significantly, variability in the ratings was reduced. Using only an iterative procedure and no normative data, Norcini, Lipner, Langdon, and Strecker (1987) also found a reduction in standard variability. No research could be found that examined the advantages and disadvantages of an iterative approach only, normative data only, or a combination.

The method is appropriate only for dichotomously scored items (Pulakos, Wise, Arabian, Heon, & Delaplane, 1989). However, the method could be modified for tests composed of continuous scale measures (e.g., assessments based on rating scales) by asking subject matter experts (SMEs) to estimate the most likely, or average, rating for minimally competent persons. Averaging these ratings across the performance measures provides a cutoff recommendation for each SME. SME recommendations can then be averaged to give an overall test standard.

Ebel method. The Ebel method requires SMEs to classify test items on two dimensions: (a) difficulty and (b) relevance. Ebel suggested three levels of difficulty (easy, medium, and hard) and four levels of relevance (essential, important, acceptable, and questionable). However, the dimensions and number of levels can be changed without altering the basic method. After considering each item on the two dimensions, SMEs working independently allocate each item to 1 of the 12 cells formed by the 3 (difficulty) x 4 (relevance) matrix. For example, item 1 might be judged to be "easy" and of "questionable relevance"; item 2 might be judged to be "hard" and "essential"; etc. Working as a group, SMEs then decide the percentage of minimally competent examinees who would be able to correctly answer items in each of the 12 cells. Percentages are assigned to cells without regard to the particular items in each cell. For example, 90% of minimally competent examinees might be expected to correctly answer "easy and essential" items; 20% might be expected to correctly answer "hard and questionable" items; etc.

For each SME, the number of items in a particular cell is multiplied by the percentage assigned to that cell. These products are summed across cells to yield a cutoff score for each SME. The average cutoff score across SMEs becomes the test standard.

SMEs find the traditional Ebel method easy to understand and implement. However, it is time-consuming. Boredom and fatigue may become a problem, especially if the test contains many items, and setting multiple cutoff scores exacerbates the problem. Other disadvantages are associated with the method. Poggio (1984) found that many SMEs were troubled by the "questionable" label on the relevance dimension. Because the dimensions and number of levels within a dimension are irrelevant to the basic method, this problem can easily be eliminated. Another disadvantage is that the Ebel method consistently results in stricter standards than other standard setting methods (Andrew & Hecht, 1976; Poggio, 1984, Skakun & Kling, 1980).

Unmodified, the Ebel method is restricted to use with dichotomously scored items (Pulakos et al., 1989). Similar to the Angoff method, the Ebel method could be modified for tests composed of items measured on a continuous scale. The original Ebel questions essentially ask for the average score (i.e., percent passed) of minimally competent persons. A modified version would be to ask SMEs to estimate the most likely rating, or average rating, for the measures within each of the matrix cells. Averaging these ratings across the cells, weighted for the number of measures in each cell, provides a cutoff recommendation for each SME. SME recommendations can then be averaged to give an overall test standard.

In the suggested modifications, the Ebel method differs from the Angoff method only in that SMEs rate categories of items instead of individual items, thus, the Ebel method requires that items be categorized. The modifications suggest the same question for both methods: What is the average score for minimally competent persons?

Jaeger method. Poggio (1984) points out that many raters have difficulty determining the percentage of examinees who should correctly answer each item. The Jaeger method circumvents that problem by having raters answer a yes/no question. Instead of trying to estimate the performance of minimally competent individuals, judges are asked to consider the following question: "Should every examinee in the population of those who receive favorable action on the decision that underlies use of the test be able to answer the test item correctly?" (Jaeger & Keller-McNulty, 1986, p. 14). In other words, should every person who is at least a minimally competent examinee be able to answer this item correctly? A "yes" response is scored as 1, and a "no" response is scored as 0.

In the first phase, judges independently answer the above question for each test item. An initial cutoff score for each judge is calculated by summing his or her "yes" responses across items. An initial test standard is determined by computing the median cutoff score across judges. While the Nedelsky and Angoff methods have been modified to include the use of normative data and an iterative approach (Cross, Impara, Frary, & Jaeger, 1984; Koffler, 1980; Norcini et al., 1987), the Jaeger method prescribes these conditions at a minimum.

For the iterative approach, the percentage of examinees who actually answered each item correctly on a recent administration of the test is presented after SMEs make their initial judgments. Upon reviewing the data, judges are asked to reconsider their recommendations and again independently answer the same question for each item. A second cutoff score is computed for each judge, and a second test standard is computed for the entire group.

In preparation for the final rating phase, more normative data is provided. Specifically, given the group's second standard, judges are told the percentage of examinees who would have failed the test on a recent administration. The distribution of cutoff scores recommended by fellow judges during the second phase is also presented. Judges once again answer the same yes/no question. Using the same computational procedure, a final standard is calculated for each judge and for the group. The median standard for the group becomes the test standard.

The Jaeger method inherently requires that judgments be made in a workshop setting. The nature of the information presented and the ensuing discussion requires a skilled workshop leader. The advantages and disadvantages of a workshop setting depend upon the frequency with which standards are set and the standard setting experience of the raters. If standards are to be set frequently as with an ongoing minimum competency testing program, a workshop approach will quickly become expensive and time-consuming.

Perhaps the greatest disadvantage of the Jaeger method is that, like the traditional Ebel method, it is time-consuming. While fatigue and boredom may become a problem, it is not likely to be as pervasive as with the Ebel method.

The Ebel method requires judges to consider items on two dimensions and little time is allotted to group discussion. Although raters answer the same question several times with the Jaeger method, only a simple yes or no answer is required, and more time is allotted to group discussion.

Finally, like the previously discussed methods, the Jaeger approach was also designed for use with dichotomously scored items. And like the other methods, the basic question put to the SMEs can be restructured to adapt the method to tests composed of continuously scored scales. In this case, the appropriate question could be stated as follows: What is the lowest score that should be observed among persons who receive favorable actions on the decision that underlies use of the test? Or simply, what is the lowest acceptable score?

The Jaeger method can be viewed as a combination of the item-based and examinee-based approaches to standard setting. Item-based approaches require decisions about test items, and examinee-based approaches require decisions about examinees. By using normative data, the Jaeger method requires decisions about test items in light of examinee performance on those items. Examinee-based methods are discussed next.

Examinee-Based Methods

A basic assumption underlying examinee-based approaches is that judges who are familiar with examinee performance in the knowledge, skill, and ability (KSA) being tested are capable of identifying individuals who are high in the KSA and those who are low. In other words, it is assumed that expert judges can conceptualize distinct levels of performance, and independent of data from the test in question, can identify individuals at each level. Not only are experts quite accurate in predicting the performance of individuals whom they know well, but also laypeople feel confident in those predictions. For example, in education where teachers serve as standard setting judges, parents readily accept the standards established via an examinee-based approach (Poggio, 1984). They often feel minimum competency testing is unnecessary because teachers can identify competent and non-competent students without using the test data.

Evaluations required of supervisors are similar to those required of teachers. In addition to formal evaluations, supervisors make informal assessments of subordinates who need remedial training, of those who are ready for additional responsibility, etc. In order to make these assessments, supervisors must be familiar with the KSAs required by the job as well as the performance of subordinates in regard to those KSAs. Furthermore, most of these assessments-are made without the aid of test data.

A second assumption underlying examinee-based standard setting approaches is that most judges are more accustomed to making decisions about individuals than making decisions about test items. This is especially true of supervisors. As mentioned earlier, most supervisory decisions are made without relying on test data. In fact, supervisors rarely, if ever, administer formal tests to their subordinates. Therefore, supervisors are even more likely than teachers to be more comfortable making decisions about individuals than about test items or tests. The two examinee-based approaches discussed in the following section are (a) Contrasting Groups method and (b) Borderline Group method.

Contrasting Groups method. According to the Contrasting Groups method, judges are asked to identify individuals who fall into one of two groups: competent vs. non-competent. Once the groups have been identified, the test is administered to them, and the distributions of scores are compared. The cutoff score is selected to maximally differentiate between the score distributions of the groups. The use of two groups results in a single test standard; however, two or more standards may be set by increasing the number of groups. For example if two standards are desired, individuals may be classified as competent, marginal, or non-competent.

One drawback of the Contrasting Groups method is the subjective process of identifying competent and non-competent individuals. To eliminate the subjective judgement, a modification of the Contrasting Groups method, administering the test to instructed and non-instructed individuals, has been suggested. The assumption is that instructed individuals are competent and non-instructed individuals are non-competent. In this way, one omits the judgmental process of identifying competent and non-competent individuals.

Several methods for analyzing the distributions and selecting a standard have been proposed. The simplest method is to plot the distributions on a single graph. The score at which the distributions intersect is selected as the test standard. This method is applicable if the distributions are not coincident and if they overlap, especially if they overlap at a single, clear point. In reality, such a pattern rarely occurs. Fisher (as cited in Poggio et al., 1981) suggests several variations of statistical procedures for establishing standards which consider the shapes and relative variances of the distributions. If the groups have normal distributions and equal variances, the Linear Discriminant Function (LDF) is appropriate. If the distributions are normal and variances are unequal, the Quadratic Discriminant Function (QDF) is used. When the distributions are not normal, non-parametric analogs to the LDF and QDF (for equal and unequal variances, respectively) are appropriate.

Borderline Group method. In implementing the Borderline Group method, judges are asked to identify individuals who are "borderline" between competent and non-competent (i.e., they cannot be clearly identified as competent or non-competent). The test is administered to these individuals, and the resulting median test score for the group defines the test standard. Many of the advantages and disadvantages of the Contrasting Groups and Borderline Group methods are the same or very similar. Therefore, pros and cons of the two methods are discussed together.

To obtain accurate, unbiased standards with the Contrasting Groups and Borderline Group methods, it is imperative that individuals selected for testing be carefully identified and classified. Raters must consider only the KSAs covered by the test and classify individuals accordingly. For example, if the test covers reading comprehension, raters must classify individuals as competent, borderline, or non-competent on reading comprehension, not some other ability. Thus, raters must be familiar with the performance of the individuals being classified. However, as the raters' familiarity with the individuals being classified increases, so does the probability of halo error. Fortunately, numerous studies (e.g., Borman, 1975; Latham, Wexley, & Pursell, 1975; Pulakos & Borman, 1985) have shown that rater training is effective in reducing halo error.

SMEs report few problems identifying competent and non-competent individuals, but many have difficulty identifying "borderline" individuals (Mills, 1983; Poggio, 1984). Mills concludes that examinees may be classified as "borderline" merely because SMEs lack sufficient information on which to base a decision.

One potential disadvantage of the examinee-based methods is that the cost of administering the test as a prerequisite for setting standards may not be feasible, especially for performance tests. While administering a written test may not be very expensive, performance tests often require more resources. It is possible to circumvent a separate, and potentially expensive, test administration by using data from the first test administration to establish performance standards. In this case, standards are not known prior to the first test administration. While this may be a less expensive solution, it is difficult to convince laypeople of the credibility of standards set in this fashion.

While examinee-based methods prescribe the classification of examinees prior to test administration, Cantor (1989) applied both the Contrasting Groups and Borderline Group procedures to archival data. Although the purpose of the study was to evaluate a previously established Ebel-derived standard, it is of interest because it is the only study to use examinee-based procedures to establish standards on archival data. Several criteria that were external to the test in question were identified and used to classify examinees as competent or non-competent. Although some classification errors resulted from partial information used to classify examinees, the methodology provides a less subjective means of classifying competent and non-competent examinees.

Aside from simplicity (Poggio, 1984), the primary advantage of the Contrasting Groups and Borderline Group methods is that they are more objective than the item-based methods. Once SMEs have identified competent, borderline, and non-competent individuals, the subjective phase is complete. For the Contrasting Groups method, decisions must be made regarding the proper use of statistics, but the characteristics of the score distributions will dictate the appropriate statistical analyses to be performed. While these methods are considered more objective by many researchers and practitioners, people who do not understand the statistical manipulations may be confused and doubt the validity of standards established through their use (Poggio, 1984).

Comparisons of Item-based and Examinee-based Methods

Studies have been conducted to examine similarities and differences among methods of setting standards on written tests. However, any similarities and/or differences among standard setting procedures as applied to performance tests remain unknown. It is generally assumed that results obtained from comparisons of written tests are applicable to performance tests. Aside from some general considerations, the consensus in the literature seems to be that the process itself is not as important as whether the standards are realistic (Buck, 1977) and whether the procedure is feasible given situational constraints such as financial and human resources, time available, appropriateness of the method for the type of test being studied, etc. (Hambleton, 1980).

In most research comparing standard setting methods, only item-based procedures are examined. Most such comparisons consider the Nedelsky method and one or more additional item-based procedures. Research results have consistently shown that the Nedelsky method produces the lowest and most unreliable standards (Brennan & Lockwood, 1980; Cross et al., 1984; Halpin & Halpin, 1987; Halpin, Sigmon, & Halpin, 1983). The Ebel method tends to produce the strictest standards (Poggio et al., 1981), and the standard produced may (Halpin & Halpin; Poggio et al.) or may not (Andrew & Hecht, 1976; Poggio, 1984) be highly reliable. The Angoff and Jaeger methods produce standards that typically fall somewhere between those produced by the Nedelsky and Ebel methods with a tendency for Jaeger standards to be stricter than Angoff standards (Cross et al.; Jaeger & Keller-McNulty, 1986).

Few studies investigate examinee-based methods. In summarizing his findings across several years of standard setting for Kansas competency tests, Poggio (1984) found that standards produced by the Contrasting Groups and Borderline Group methods tend to be lower than those produced by the Angoff procedure. With one group of raters using different procedures, Mills (1983) found no differences in the standards set with the Angoff, Contrasting Groups, and Borderline Groups methods. Mills points out that although different methods may have produced different results, at least some of the discrepancies between methods probably have been due to differences between groups of judges.

In comparing the ease of implementation among methods, Poggio (1984) found the Angoff, Ebel, Contrasting Groups, and Borderline Group methods easily implementable and comprehensible. His research did not examine the Jaeger method, but he found that judges were confused by the Nedelsky method. In general, results taken across studies show that the Angoff method is the easiest to implement and that raters more readily comprehend the task they are to perform compared to other methods. Although previous research found that raters sometimes had difficulty identifying borderline individuals (Mills, 1983; Poggio et al., 1981), it is believed that an exhaustive definition including hypothetical examples can overcome this confusion.

When generalized across studies and standard setting procedures, the perception is that: (a) the Ebel method produces the highest standards, (b) the Nedelsky method produces the lowest, (c) the Angoff and Jaeger methods produce standards somewhere in the middle, and (d) most examinee-based methods are not feasible. Because of the disparity in standards established by the various procedures, many researchers recommend the use of several standard setting procedures to set performance standards (Halpin et al., 1983; Koffler, 1980).

Standard Setting Process

There are several issues concerning the standard setting process that are independent of the procedure used. Before deciding on the appropriate method, these issues warrant examination. They are discussed in the following sections.

Characteristics of judges. The identification and utilization of qualified experts is perhaps the most important consideration in any standard setting procedure. Research results in the field of education indicate that different groups of judges from a variety of backgrounds, if qualified, provide similar standards. In addition, standards are more readily accepted if they are set by qualified judges from a number of backgrounds (Andrew & Hecht, 1976; Jaeger, 1976).

Employing a variety of judgmental standard setting procedures, the U.S. Army's Synthetic Validity Project (Peterson, Owens-Kurtz, Hoffman, Arabian, & Whetzel, 1989) used Non-commissioned Officers (NCOs) and Officers from operational (Forces Command [FORSCOM]) and training (Training and Doctrine Command [TRADOC]) commands in an attempt to survey experts with a variety of experiences. While Officers had slightly more reliable ratings, there were no other appreciable NCO/Officer or FORSCOM/TRADOC differences. Thus, using an item-based method, either NCOs or Officers from FORSCOM or TRADOC could be used. Restricting the diversity of SMEs, however, raises the issue of standard acceptability. If the test and resulting standards were to be used at both FORSCOM and TRADOC sites, it would be prudent to survey SMEs from both commands.

Because both NCOs and Officers are affected by scores from job performance measures, it is advisable to use both in standard setting exercises. One could also argue that because airmen are affected by the standards, their judgments (i.e., incumbents' judgments) should be considered when defining those standards. The central issue here may be summarized by the question: Who are the users of the research results, and are they represented?

While it is important to survey SMEs from a variety of backgrounds, only SMEs who are directly familiar with airman performance are appropriate for examinee-based standard setting procedures. In most cases, NCOs work directly with airmen and consequently are more familiar with an individual airman's performance than are Officers. If an examinee-based method is used, it might be wise to use only NCO raters.

Number of judges. In addition to obtaining SMEs from diverse experiences, one must decide on the optimal number of judges. The optimal number of judges is determined to some extent by psychometric considerations, by the standard setting method employed, and by the number of qualified SMEs available. The number of judges is positively correlated with the reliability of the standard and negatively correlated with the amount of dispersion in the standard (Pulakos et al., 1989). Jaeger and Keller-McNulty (1986) suggest determining the necessary number of SMEs based on reductions of the standard error of the test standard and the standard error of measurement of the test.

Cross et al. (1984) and Jaeger and Busch (1984) found that psychometric considerations are maximized with sample sizes of 20 to 30.

One must also consider the various types of raters being surveyed (e.g., NCOs and Officers from Site A and Site B). If main effects or interactions exist for rater type, a large number of raters is needed (e.g., 20 to 30 of each rater type). Thus, for the four types of raters suggested -- Site A NCOs, Site A Officers, Site B NCOs, and Site B Officers -- a total of 80 to 120 raters would be needed. Data from the U.S. Army's Synthetic Validity Project (Peterson et al., 1989), however, indicate that such a large number of raters is unnecessary. Furthermore, much standard setting research has been conducted with as few as five to eight raters (Andrew & Hecht, 1976; Brennan & Lockwood, 1980; Plake & Melican, 1989; Skakun & Kling, 1980).

The standard setting method often imposes practical constraints when determining the optimal number of SMEs. Methods implementing group discussions necessitate small- to medium-sized groups to prevent a few dominant SMEs from exerting too much control over other judges' decisions while still providing an adequate number of divergent opinions. Workshops with 20 participants are practical, but workshops involving more than 20 participants tend to be unmanageable.

Iterative process. As previously stated, the Jaeger method is the only standard setting procedure that prescribes an iterative process. All item-based methods, however, have been modified to include an iterative process. The primary purpose of the iterative process is to provide SMEs with an opportunity to reconsider their initial cutoff scores in light of potential consequences of those scores. The iterative process tends to follow one of several formats: (a) a presentation and individual consideration of normative data, (b) a presentation of the group's standard, (c) a group discussion allowing judges to debate the rationale underlying their cutoff scores, and (d) various combinations of (a), (b), and (c). The presentation of normative data does not require an iterative process. For example, normative data can be presented in the initial phase followed by an iterative process with a group discussion (Peterson et al., 1989). A group discussion, on the other hand, does necessitate an iterative process. The following paragraphs focus on the group discussion as part of the iterative process.

A group discussion has been shown to reduce the variability in standards without significantly altering the standards (Jaeger & Busch, 1984; Norcini et al., 1987). By reducing the variability in standards, a group discussion thereby produces a more reliable standard. A few words of caution regarding the implementation of a group discussion iterative process, however, are in order. Pulakos et al. (1989) point out that individual judges' cutoffs, stated without justification, "can lead to a shift in judgment toward the central tendency of the group" (p. 29). To effectively evaluate differences in individual cutoff scores, the discussion must provide an opportunity to examine the rationale behind those cutoff scores. As in any discussion, a few dominant individuals are likely to unduly influence the group if not restrained. Therefore, a consensus discussion, the goal of which is to reach a general agreement among participants, is recommended rather than a convergent discussion, which requires unanimous agreement among participants.

In addition, a skilled workshop leader is needed to maintain a controlled discussion.

In addition to a group discussion, the iterative process often includes the evaluation of normative data collected for the test in question. Some reviewers recommend that judges review normative data when setting performance standards (Hambleton, 1978; Shephard, 1980). By examining normative data, judges can evaluate the consequences of their recommended standard. Furthermore, the use of normative data has been shown to reduce the variability in standards (Cross et al., 1984; Jaeger & Busch, 1984).

Hambleton and Powell (as cited in Pulakos et al., 1989) argue that the decision to use normative data should depend on the goals and constraints of the testing program. If the goal is to normally distribute examinees in terms of test scores, then emphasis on normative data is appropriate. However, if cutoff scores are to be used for selection purposes, too much emphasis on normative data is clearly inappropriate. A strong focus on normative data shifts the standard setting emphasis from "what performance should be" to "what performance is." In the present situation, cutoff scores are to be linked to selection standards. Therefore, emphasizing normative data would be a mistake; normative data should be used as a reality check only (i.e., to demonstrate the consequences of the established cutoff scores).

Training of judges. For the consideration of normative data to be effective, judges must be trained in their use. One cannot assume that judges can properly interpret even the "simplest" types of normative data (e.g., frequency distributions). SMEs must be taught how to properly read and interpret frequency distributions, graphs, etc. If more complex data are to be used (e.g., estimated item difficulty values), the meaning of these data must be carefully explained.

The second aspect of judge training involves insuring that SMEs fully understand the task they are to perform and familiarizing them with the test on which they will be setting standards. Explaining the standard setting procedure to be used may be fairly straightforward depending on the method being used. Clear, concise workshop instructions may be all that are necessary to ensure that judges understand the task at hand. On the other hand, Norcini et al. (1987) included a practice session in their explanation of the standard setting procedure being used. To familiarize SMEs with the test under consideration, Cross et al. (1984) and Jaeger and Busch (1984) had judges actually complete the test under approximately normal test administration conditions. At the very least, the instructions should include information about the way the test was administered and scored.

If an examinee-based method is used, training in the avoidance of halo error should also be conducted. As previously stated, examinee-based methods require the use of judges who are extremely familiar with the KSAs covered by the test as well as the performance of the individuals being classified in regard to those KSAs. Also, there is a positive relationship between rater-ratee familiarity and halo error (i.e., the more familiar the rater is with the ratee the more likely he or she is to commit halo error). Pulakos and Borman (1985), and many other researchers, have developed a rater training program which has been shown to reduce rating error. Thus, some sort of rater training program should be conducted if an examinee-based method is used.

Number of standards. An additional issue concerns the number of standards desired. Is a single pass/fail score appropriate, or would several levels of performance standards be more beneficial? In many testing situations, several levels of performance are defined with performance below a certain point deemed unacceptable. In education for example, 90% correct or greater is often regarded as outstanding, 80% to 89% correct is superior, 70% to 79% correct is acceptable, and 69% correct or below is unacceptable. Although not explicitly stated, the purpose of various levels of performance is to encourage individuals to strive for improvement. Because the goal in war is to be better skilled than the enemy, airmen should never be encouraged to "rest on their laurels" once they have met the minimum performance standard. In maintaining job performance skills, the goal should be to strive for perfection. For these reasons, it is suggested that standards be established to differentiate among several levels of performance (e.g., unqualified, qualified, superior, and distinguished).

Performance standard definitions. A final consideration is the performance definition against which standards will be set. Performance definitions, in concept, determine what it means for an airman to be distinguished, superior, qualified, or unqualified. The question is whether the definition should be provided by researchers or by the SMEs. While no research could be found to demonstrate the superiority of either researcher- or rater-generated performance definitions, it seems prudent to begin the session with performance defined by the researcher. If the definition is completely out of line, raters can enhance it with the guidance of the researcher. If more than one workshop is to be conducted, the definition can be corrected at the first workshop. The corrected definition can then be used in subsequent workshops.

Setting Job Performance Standards for the JPM Hands-on Tests

After careful consideration of the standard setting literature reviewed above, two methods of establishing minimum performance standards were proposed for setting standards on the AGE hands-on tests (Fotouhi, Mosher, & McCloy, 1990). One method requires the use of SME judgments and was termed the item-based judgmental technique. The second method uses only the JPM database and was termed the examinee-based archival technique. As noted above, there are numerous advantages to identifying multiple levels of competency. Therefore, both procedures sought to establish five levels of job proficiency. The proficiency levels correspond to the 5-point rating scale developed for the JPM Project. The proficiency levels and their behavioral definitions are presented in Figure 1.

The judgmental and archival standard setting procedures are described in detail in the following sections. The judgmental technique is presented first followed by the archival technique and then a comparison of the two methods.

Unqualified Does not display knuwledge and skill necessary to properly complete tasks and assignments; unable to perform without direct supervision; often fails to complete assignments; performs more slowly than other first-term airmen.

Marginal Occasionally displays adequate knowledge about how to complete tasks and assignments; quality of work is inconsistent; requires direct supervision on most tasks to ensure quality and accuracy; usually completes tasks within required time.

Qualified Displays good knowledge/skill in most aspects of the job; able to properly complete the majority of tasks; requires supervision only on difficult tasks and assignments; completes work in the same time as other first-term airmen.

Distinguished Displays considerable knowledge and skill to complete assignments and tasks properly; performs effectively with little supervision; completes tasks more quickly than the average first-term airman.

Exceptional Displays exceptional knowledge/skill to consistently complete assignments and tasks properly; requires little or no supervision; completes tasks in minimum time.

Figure 1. Proficiency levels for setting job performance standards.

Item-Based Judgmental Technique

The item-based judgmental technique used in the present study is based on the Jaeger standard setting paradigm (Jaeger & Keller-McNulty, 1986). Two significant modifications to the Jaeger paradigm were made. First, as originally proposed, the Jaeger method requires judges to examine and make decisions on every test item. Item decisions are then aggregated to yield a standard for the test. As mentioned previously, five levels of competency were to be established on 16 hands-on tasks which range in length from 7 to 30 items. It was decided that an examination of every item would be prohibitively time consuming; therefore, judges were asked to set standards for each task.

The second modification reduced the number of iterations to two, down from Jaeger's proposed three iterations. As originally proposed by Jaeger, SMEs first set standards by examining test items only. Before the second iteration, normative data from a recent administration of the test are presented for the judges' consideration. Prior to the third iteration, additional normative data are presented. Research has demonstrated that the examination of normative data reduces variability in standards (Jaeger & Busch, 1984). Because time was limited, normative data were presented in the initial phase, thereby combining Jaeger's first and second iterations. Related to the presentation of normative data, the Jaeger method includes a consensus discussion between iterations. During the discussion, judges share their rationales for divergent cutoffs with the intent of reducing variability in standards. Also, the ramifications of implementing the judges' standards are explained. By reducing the number of iterations to two, the present method employed only one consensus discussion as opposed to Jaeger's originally proposed two.

With these modifications, the judgmental procedure was as follows. The initial procedures were pilot tested with a small group of SMEs at Tactical Air Command (TAC) Headquarters, Langley AFB. The primary goal of the pilot test was not to collect data but to ensure that the instructions and materials would be easily understood by the target audience of senior AGE NCOs. Five senior NCOs from the AGE AFS participated in the pilot test. As they set standards in compliance with our instructions, the NCOs recommended modifications that would clarify the standard setting process. Their recommendations were reasonable and easily implemented, and the standard setting procedure discussed below reflects their suggested modifications.

Method

Participants

Fifteen NCOs from two Tactical Training Wings (TTWs) served as SMEs in the data collection: nine at Tyndall AFB and six at MacDill AFB. SMEs had an average of 10.7 (SD = 4.9) years of experience in the AGE AFS and thusly were familiar with the AGE career field and had 12.1 (SD = 4.7) years experience in the Air Force. Of the 15 participants, 2 held the rank of E-4 or Sergeant, 9 of E-5 or Staff Sergeant, 3 of E-6 or Technical Sergeant, and 1 of E-8 or Senior Master Sergeant. As indicated by their rank, most SMEs had management experience and, on the average, supervised five subordinates.

Procedure

After the pilot test at Langley AFB to finalize procedures and forms, two four-hour workshops were held: one at Tyndall AFB and one at MacDill AFB. SMEs first received a briefing which outlined the goals of the project. During the briefing, workshop leaders also clarified any questions SMEs wanted to ask. After the briefing and the ensuing question and answer session, participants completed a Background Information Form. In addition to identifier and experience questions, the Background Information Form contained the Privacy Act statement which all participants were directed to read and sign.

In setting minimum job performance standards, SMEs were told to concentrate on the airman with 28 months time in service including basic and technical training. The 28-month airman was selected as the prototype because the average AGE incumbent in the JPM project had been in the Air Forcc for 28 months. The performance levels and definitions (see Figure 1) were explained. Participants accepted the levels and definitions and indicated that they could identify airmen who fell into each category.

To familiarize themselves with the tests for which standards were being set, SMEs reviewed hands-on test summaries. For test security reasons and because of time constraints, actual copies of the hands-on tests were not used. However, the test summary format closely resembled the actual test in that it presented the test station setup, verbal instructions, and the steps to be performed in an outline format. Figure 2 presents a copy of the test summary for Test 154--Perform an Aircraft Support Generator Service Inspection. During the practice test, SMEs were confused about how the hands-on tests were used. SMEs tended to think in terms of a written test and thought that the incumbent was given a written copy of the test and expected to perform the steps listed. To prevent confusion in the data collection workshops, the workshop leaders repeatedly explained that the test was used by the test administrator only and was never seen by the incumbent. SMEs were instructed to think of the summaries as a score sheet rather than a test and to equate hands-on test administration with over-the-shoulder Quality Assurance evaluations.

Workshop leaders then presented the standard setting form which included actual hands-on test score information from the JPM Project. Test score information was presented in four columns. The first column presented test scores at .25 test score intervals. The subsequent columns presented the percentage of airmen who received the adjacent score as well as the percentage who scored above and below the adjacent score. Figure 3 presents a copy of the standard setting form for Test 154--Perform an Aircraft Support Generator Service Inspection. Workshop leaders explained how to interpret the normative data and how to use it to guide standard setting decisions.

After carefully reviewing the proficiency levels, their behavioral definitions, and the test summary, SMEs independently indicated the <u>minimum</u> score an airman could obtain to be considered Marginal, Qualified, Distinguished, and Exceptional by drawing a line below the score that represents the lowest score corresponding to each of the four performance levels. They labeled their lines "M" for Marginal, "Q" for Qualified, "D" for Distinguished, and "E" for Exceptional. (Note that any score below the minimum Marginal score denotes Unqualified performance.)

We anticipated that SMEs would need to practice setting standards on two or three tests before becoming familiar with the procedure. Once comfortable, we thought SMEs would be able to quickly set standards on the remaining tests. The pilot test confirmed our hypothesis. Therefore, the standard setting process required SMEs to independently set standards one test at a time for three tests, but they worked at a group pace. That is, one test was distributed, and SMEs independently set standards on that test. When the group finished setting standards on the first test, the second test was distributed. After setting standards in this manner on three tests, SMEs were given two booklets: one containing the test summaries, and one containing the corresponding standard setting forms. The booklets allowed SMEs to proceed at

Perform an Aircraft Support Generator Service Inspection
(Look Phase Only--Maintenance or Servicing Deferred)

Tools & Equipment: Applicable T.O., common screwdriver

Configuration: -86 in work bay with access panels opened/removed. The maintenance forms should reflect no delayed discrepancies on the unit that would affect the inspection.

Instructions: *This task requires you to perform a service inspection on this generator. You may use any technical data normally available to you for the performance of this task. If any servicing is required. Tell me about it but do not perform the servicing. If you discover any discrepancies during the inspection, tell me what they are but do not correct them or annotate the forms.*

1. Check the maintenance forms to determine current equipment status?
2. Check the fuel level?
3. Check the engine oil level?
4. Visibly check the air inlet screen for blockage?
5. Check the operation of the hand brake by setting the brake and nudging the unit to insure that the brake holds?
6. Visually check the tires for deflation and damage?
7. Visually check the output cables for damage, worn insulation, condition of pins, pins extending beyond the cable head insulation, and other discrepancies?
8. Check the air filter service indicator for a red indication?
9. Visually check for fuel, oil or coolant leaks (3 drops per minute for oil and coolant, none for fuel)?
10. Visually check the enclosure and chassis for serviceability?
11. Visually check the external power receptacle for serviceability?
12. Visually check gauges for serviceability?
13. Ensure serviceability of lamps?
14. Assure the operation of the emergency shut down lever?
15. Inspect the drive belts for proper tension, fraying or damage?

Figure 2. Hands-on test summary for Test 154--Perform an Aircraft Support Generator Service Inspection.

their own pace. This strategy allowed SMEs to master the procedure without being overwhelmed by the size of the assignment.

Time did not allow judges to practice the standard setting process on hypothetical tests. Therefore, three tests were selected from among the 16 hands-on tests to serve as "practice tests." Several factors were considered in selecting practice tests. Primary emphasis was placed on identifying tests that might cause concern among the SMEs. Based on the pilot test at Langley AFB, tests on which airmen scored poorly caused the most confusion among SMEs. Tests with long or complicated test summaries were also considered as candidates for practice tests. It was decided that the first practice test

Perform an Aircraft Support Generator Service Inspection
(Look Phase Only--Maintenance or Servicing Deferred)

Test Score	% who Received this Score	% who Scored Worse than this	% who Scored Better than this
10.00	13%	87%	0%
9.75	0%	87%	13%
9.50	0%	87%	13%
9.25	20%	67%	13%
9.00	0%	67%	33%
8.75	21%	46%	33%
8.50	0%	46%	54%
8.25	0%	46%	54%
8.00	15%	31%	54%
7.75	0%	31%	69%
7.50	0%	31%	69%
7.25	11%	20%	69%
7.00	0%	20%	80%
6.75	8%	12%	80%
6.50	0%	12%	88%
6.25	0%	12%	88%
6.00	3%	9%	88%
5.75	0%	9%	91%
5.50	0%	9%	91%
5.25	3%	6%	91%
5.00	0%	6%	94%
4.75	3%	3%	94%
4.50	0%	3%	97%
4.25	0%	3%	97%
4.00	2%	1%	97%
3.75	0%	1%	99%
3.50	0%	1%	99%
3.25	0%	1%	99%
3.00	0%	1%	99%
2.75	1%	0%	99%
2.50	0%	0%	100%
2.25	0%	0%	100%
2.00	0%	0%	100%
1.75	0%	0%	100%
1.50	0%	0%	100%
1.25	0%	0%	100%
1.00	0%	0%	100%
0.75	0%	0%	100%
0.50	0%	0%	100%
0.25	0%	0%	100%
0.00	0%	0%	100%

Figure 3. Standard setting form for Test 154--Perform an Aircraft Support Generator Service Inspection.

should be fairly "easy" in terms of meeting SME expectations of an acceptable score distribution and in terms of the test summary. It was decided that the second and third practice tests should be progressively more "difficult" in terms of score distribution expectations and test summaries. The order in which the practice tests were presented was (a) Test 238--Splice Electrical Systems Wiring, (b) Test 503--Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance Information, and (c) Test 179--Perform a Gas Turbine Compressor Periodic Inspection. The test summary and standard setting form booklets presented the remaining tests in numerical order by their code numbers.

After setting their initial standards, SMEs were given a break while the workshop leaders tallied the group's first round cutoffs. After the break, the workshop leaders guided the group in a discussion of the three tests with the most discrepant cutoffs. Due to time constraints, statistical measures of discrepancy (i.e., variance) could not be calculated; therefore, discrepant cutoffs were identified by a cursory inspection of the tallies. During the discussion, participants with the highest and lowest cutoff scores at each proficiency level (Marginal, Qualified, Distinguished, and Exceptional) stated the rationale for their cutoffs. Workshop leaders pointed out the ramifications of implementing these cutoffs by noting the percentage of incumbents that would be classified at, below, and/or above the highest, lowest, and midpoint cutoffs for each proficiency level. Although time constraints permitted a thorough discussion of only three tests, the workshop leaders summarized the rationales and described how those rationales could be generalized to other tests.

Immediately following the discussion, SMEs repeated the standard setting process. During the final phase, SMEs once again set standards independently. SMEs proceeded at their own pace using test summary and standard setting form booklets which presented the tests in the same order as in the initial session.

Results and Discussion

Qualitative feedback. Table 2 presents the tests discussed at Tyndall AFB and at Langley AFB in the order in which they were discussed and the range of cutoffs for each proficiency level. At each base, the test with what appeared to be the most discrepant cutoffs was discussed first.

Based on comments from the discussion sessions, three general strategies appeared to guide judges' standard setting decisions. The strategies are:

1. Criticality. A criticality strategy was used to set strict standards on life-threatening tests/tasks.

2. Difficulty. If a test/task was perceived as being particularly difficult, SMEs were willing to set lenient standards. Conversely, SMEs tended to set strict standards for "easy" tests/tasks.

3. Frequency of Performance. SMEs tended to set stricter standards for tests/tasks that were performed frequently compared to those performed infrequently.

Table 2

Tests Discussed at Each Base

Base	Test	Cutoff Ranges			
		Marginal	Qualified	Distinguished	Exceptional
Tyndall	179	0.25 - 3.75	1.25 - 8.25	2.25 - 9.25	3.50 - 10.00
	251	0.00 - 4.75	1.50 - 6.75	3.75 - 9.00	4.50 - 10.00
	446	1.00 - 5.50	2.00 - 7.00	4.25 - 8.75	5.00 - 9.75
MacDill	179	0.75 - 5.00	2.75 - 6.25	5.75 - 8.00	8.00 - 9.50
	155	3.00 - 7.00	6.00 - 8.00	8.00 - 9.00	9.00 - 10.00
	209	2.50 - 6.75	4.00 - 7.75	6.50 - 8.75	8.50 - 10.00

Note. Tests are:
179 = Perform a Gas Turbine Compressor Periodic Inspection
155 = Perform a Service Inspection on a Load Bank
209 = Measure Resistance in AGE Electrical Systems
251 = Adjust Turbine Engine Fuel System Components
446 = Isolate Pneumatic System Malfunction

When only one strategy applied to a test, implementing that strategy was fairly easy. However for most tests, two or even all three strategies applied. SMEs expressed frustration at reconciling various combinations of the strategies. For example, if a task was infrequently performed but was life threatening, the strategy called for lenient standards because of infrequent performance and strict standards because of criticality. When strategy compromises had to be made, SMEs seemed unable to clarify whether the strategies were weighted equally or, if not weighted equally, which strategies received the most weight.

Another problem concerned differences in the frequency of task performance across duty stations. Some of the tasks tested are performed on a regular basis at some duty stations and either infrequently or not at all at others. Differences in task performance frequency across duty stations were compounded when various combinations of standard setting strategies had to be reconciled. As with the problem of competing strategies, SMEs were never able to explain how they resolved duty station differences in frequency of task performance.

A few judges questioned the normative data. They did not believe that current, first-term airmen perform as some of the test score distributions indicate. Other judges noted that several of the tasks tested are complicated and/or are performed infrequently. As a result, first-term airmen have little or no opportunity to perform them. These judges pointed out that even 4- or 5-year airmen are not trusted to perform such tasks without supervision.

Descriptive statistics. Table 3 presents means and standard deviations for the standards set on each test for both initial ratings (Round 1) and re-ratings (Round 2). In addition, means and standard deviations across all tests for Round 1 and 2 ratings are presented at the bottom of the table. It should be noted that these means represent the proficiency level cutoffs and that the standard deviations are one index of rater agreement. Figure 4 graphically presents the cutoffs across all 16 hands-on tests for Round 1 and Round 2.

An examination of the data in Table 3 and Figure 4 suggests that neither cutoffs (i.e., means) nor rater agreement (i.e., standard deviations) change greatly from Round 1 to Round 2. Repeated measures analysis of variance (ANOVA) was used to test differences between Round 1 and 2 cutoffs and rater agreement. Analyses were conducted with the hands-on tests as cases and the statistics as the repeated "trials." In testing mean cutoff differences, significant main effects were found for round ($F_{(1,15)}$=28.038, p=.000) and cutoff ($F_{(3,45)}$=1021.059, p=.000), and the interaction was also significant ($F_{(3,45)}$=53.531, p=.000). In testing rater agreement, significant main effects were found for round ($F_{(1,15)}$=6.639, p=.021) and cutoff standard deviation ($F_{(3,45)}$=65.488, p=.000), but the interaction was not significant ($F_{(3,45)}$=0.842, ns).

Based on the ANOVA results, planned comparisons were made to determine precisely where the significant differences lie. Table 4 presents the results of planned comparisons for the Marginal, Qualified, Distinguished, and Exceptional cutoff means and standard deviations. Results indicate that the discussion and re-rating process influence the cutoffs for all levels of proficiency except Exceptional, but the process influences rater agreement for only the Qualified and Marginal proficiency levels. Closer inspection of the data at the bottom of Table 3 indicates that the discussion and re-rating process led to standards that are about .50 test score points higher at each cutoff and to a slight increase (.20) in rater agreement.

Reliability. Variance component analyses and generalizability coefficients (e.g., Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Shavelson, Webb, & Rowley, 1989) were calculated to examine reliability by round and hands-on test. Interrater reliabilities by round for each hands-on test are presented in Table 5. In addition, Table 5 presents average reliability across all 16 hands-on tests by round. Average reliability was calculated by transforming the interrater reliability coefficient for each hands-on test to a Fisher z coefficient, averaging Fisher z coefficients, and transforming the averaged Fisher z back to a reliability coefficient. Using this procedure, average interrater reliability across all hands-on tests for both Round 1 and Round 2 is .97. As confirmed by a repeated measures ANOVA, interrater reliability does not change significantly from Round 1 to Round 2 ($F_{(1,15)}$= 1.103, ns).

Conclusions

Based on comments made during the discussion, it is not surprising that cutoff and rater agreement differences between rounds were not significant at the higher proficiency levels. In terms of cutoffs, SMEs seemed to share the same opinions as to what constitutes Distinguished or Exceptional performance.

Table 3

Judgmental Technique: Minimum Test Score Cutoff by Test and Round

Test	Marginal		Qualified		Distinguished		Exceptional	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
154 (Perform Aircraft Support Generator Service Inspection)								
Mean	5.77	6.30	7.23	7.48	8.57	8.73	9.65	9.70
SD	1.42	0.72	1.03	0.49	0.71	0.41	0.38	0.30
155 (Perform Service Inspection on a Load Bank)								
Mean	4.32	4.98	6.20	6.50	7.93	7.97	9.32	9.12
SD	1.49	1.67	1.12	1.54	0.93	1.35	0.62	1.30
162 (Perform Service Inspection on Hydraulic Test Stand)								
Mean	5.15	5.63	6.75	7.00	8.38	8.42	9.53	9.52
SD	1.07	0.63	0.80	0.67	0.60	0.61	0.39	0.45
179 (Perform Gas Turbine Compressor Periodic Inspection)								
Mean	2.28	3.85	4.18	5.35	6.25	7.00	7.68	8.20
SD	1.54	2.07	1.85	1.84	1.82	1.61	1.89	1.56
209 (Measure Resistance in AGE Electrical Systems)								
Mean	4.73	5.22	6.37	6.60	8.03	8.15	9.28	9.13
SD	1.49	0.93	1.30	0.81	0.88	0.74	0.53	0.63
215 (Perform AGE Electrical Systems Operational Checks)								
Mean	3.78	4.77	5.82	6.47	7.98	8.05	9.08	9.17
SD	1.72	1.79	1.73	1.53	1.36	1.50	1.26	1.42
238 (Splice Electrical Systems Wiring)								
Mean	3.93	5.67	6.00	7.03	8.13	8.55	9.33	9.57
SD	1.45	1.07	1.18	0.94	1.08	0.60	0.69	0.33
251 (Adjust Turbine Engine Fuel System Components)								
Mean	2.85	4.35	5.05	5.92	7.28	7.83	8.82	9.17
SD	1.93	1.66	1.72	1.19	1.30	0.68	1.37	0.52
260 (Clean Motor and Generator Components)								
Mean	4.43	5.68	6.70	7.17	8.65	8.68	9.73	9.63
SD	2.09	1.20	1.51	0.96	0.61	0.72	0.24	0.50

(Table 3 continued)

Test	Marginal		Qualified		Distinguished		Exceptional	
	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2	Round 1	Round 2
264 (Isolate Engine, Motor, or Generator Malfunction)								
Mean	4.27	4.65	5.90	6.00	7.83	7.57	8.92	8.78
SD	1.39	1.04	1.23	0.95	0.85	0.97	0.88	0.93
284 (Remove and Replace Engine Fan Belts)								
Mean	6.10	6.55	7.52	7.53	8.90	8.77	9.82	9.75
SD	1.09	1.05	0.70	0.91	0.36	0.67	0.32	0.40
300 (Remove or Install Fuel Line and Fittings)								
Mean	5.72	5.85	7.02	7.18	8.47	8.57	9.55	9.57
SD	1.15	0.79	0.99	0.61	0.57	0.50	0.48	0.42
421 (Remove or Install Hydraulic Lines)								
Mean	4.50	5.22	6.08	6.50	7.70	8.03	9.02	9.07
SD	1.84	2.21	1.80	1.96	1.58	1.35	1.25	1.14
446 (Isolate Pneumatic System Malfunctions)								
Mean	3.60	4.80	5.28	6.03	7.32	7.93	8.73	9.18
SD	1.75	1.83	1.74	1.79	1.36	0.97	1.28	0.66
503 (Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance Information)								
Mean	3.75	5.72	5.58	7.08	7.68	8.30	9.17	9.37
SD	1.52	1.44	1.46	1.34	1.47	1.24	1.37	0.93
549 (Inspect Vehicles for Safety of Operation)								
Mean	5.35	6.62	6.98	7.82	8.67	9.02	9.67	9.83
SD	1.70	0.94	1.36	0.75	0.67	0.47	0.34	0.31
Across all Tests								
Mean	4.41	5.37	6.17	6.73	7.99	8.22	9.21	9.30
SD	1.83	1.55	1.59	1.36	1.24	1.07	1.06	0.91

Note. N = 15 for all cells.

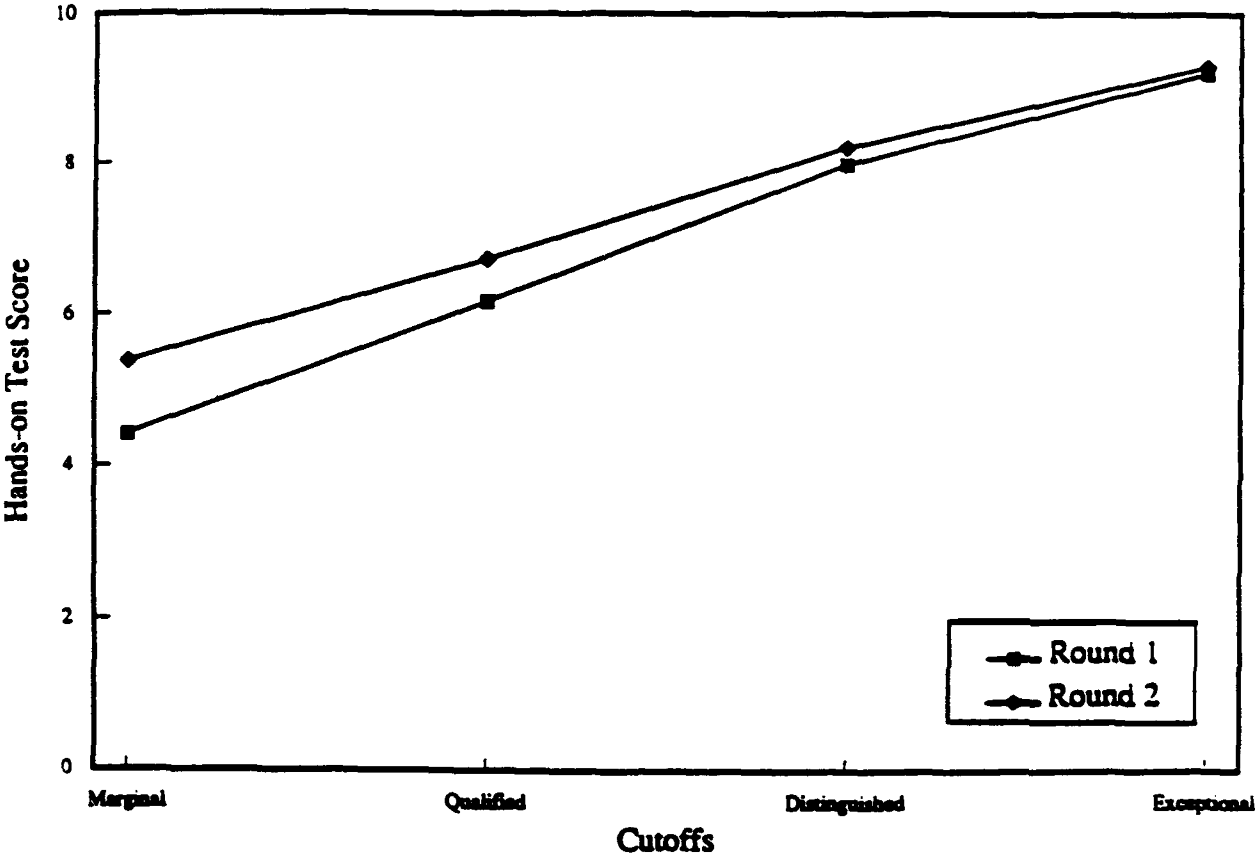

Figure 4. Judgmental technique: Standards across all hands-on tests for Round 1 and Round 2.

At the lower proficiency levels (i.e., Marginal and Qualified), judges seemed to disagree as to how leniently they wanted to treat such performers. Judges who seemed to express tolerance of subordinate errors and willingness to closely supervise their subordinates appeared to set lower Marginal and Qualified cutoffs. Judges who wanted their subordinates to be able to work without close supervision and who were less tolerant of subordinate errors seemed to set higher Marginal and Qualified cutoffs.

In support of other standard setting research (Jaeger & Keller-McNulty, 1986; Pulakos et al., 1989), variability in standards, in terms of standard deviations, decreased from Round 1 to Round 2. However, contrary to the bulk of standard setting literature (Jaeger & Keller-McNulty; Pulakos et al.) mean cutoffs *increased* between rounds. The reason for an increase in standards is unclear. It is not likely to be due to a lack of control in the discussion session. SMEs who favored strict cutoffs were not allowed to dominate the discussion. Furthermore, judges who endorsed lenient standards did not appear to be intimidated by their colleagues who supported strict standards.

The observed increase in standards may be the result of setting multiple

Table 4

Planned Comparisons for Round Effects on Cutoff Means and Standard Deviations Using Repeated Measures ANOVA

Source	SS	df	MS	F	p
Round Effects on Cutoff Means:					
Marginal Cutoff	14.665	1	14.665	48.151	0.000
Error	4.568	15	0.305		
Qualified Cutoff	5.064	1	5.064	27.623	0.000
Error	2.750	15	0.183		
Distinguished Cutoff	0.895	1	0.895	10.455	0.006
Error	1.285	15	0.086		
Exceptional Cutoff	0.132	1	0.132	2.858	0.112
Error	0.691	15	0.046		
Round Effects on Cutoff Standard Deviations:					
Marginal Cutoff	0.802	1	0.802	4.783	0.045
Error	2.516	15	0.168		
Qualified Cutoff	0.651	1	0.651	6.850	0.019
Error	1.426	15	0.095		
Distinguished Cutoff	0.203	1	0.203	2.558	0.131
Error	1.191	15	0.079		
Exceptional Cutoff	0.142	1	0.142	1.057	0.320
Error	2.014	15	0.134		

Note. Each case is represented by a hands-on test; means and standard deviations for ratings within each test are the variables.

cutoffs for the military. A decrease in standards has been found in studies which focused on obtaining a single, pass/fail cutoff for educational and certification tests (Jaeger & Keller-McNulty, 1986; Pulakos et al., 1989). In the Army's Synthetic Validity Project where multiple cutoffs were established, however, an increase in standards was found (Wise, Peterson, Hoffman, Campbell, & Arabian, 1990). In both the present study and in the Wise et al. study, some SMEs stated a preference for the military standard of perfection (i.e., 100%) as the lowest acceptable score defining Qualified performance.

Increase/decrease differences in standards may also be caused by the level at which judges set standards. To set standards for educational and certification tests, judges evaluate test *items* (Jaeger & Keller-McNulty,

Table 5

Reliability of Hands-on Tests by Round

Hands-on Test No.	Title	Reliability Round 1	Reliability Round 2
154	Perform Aircraft Support Generator Service Inspection	.98	.99
155	Perform Service Inspection on a Load Bank	.98	.96
162	Perform Service Inspection on Hydraulic Test Stand	.99	.99
179	Perform Gas Turbine Compressor Periodic Inspection	.96	.94
209	Measure Resistance in AGE Electrical Systems	.98	.99
215	Perform AGE Electrical Systems Operational Checks	.97	.96
238	Splice Electrical Systems Wiring	.98	.98
251	Adjust Turbine Engine Fuel System Components	.97	.98
260	Clean Motor and Generator Components	.98	.98
264	Isolate Engine, Motor, or Generator Malfunctions	.98	.98
284	Remove and Replace Engine Fan Belts	.99	.98
300	Remove or Install Fuel Line and Fittings	.98	.99
421	Remove or Install Hydraulic Line	.95	.93
446	Isolate Pneumatic System Malfunctions	.97	.96
503	Research T.O.s for Information	.97	.96
549	Inspect Vehicles for Safety of Operation	.97	.98
Across all Hands-on Tests		.97	.97

Note. Number of raters for each hands-on test equals 15.

1986; Pulakos et al., 1989). In the present study, SMEs set standards at the test level; and judges in the Wise et al. (1990) study set standards on broad dimensions of job performance (e.g., Individual Combat, Communications). Although SMEs in the present study set test standards after reviewing test items, they may have focused more on the test title (i.e., the task) when indicating their cutoffs. By focusing on the test title, their standards may have been influenced primarily by task characteristics (i.e., criticality, frequency of performance, etc.); SMEs may have not known how to use item information to guide cutoff decisions.

In summary, the Round 2 increase in cutoffs found in the present study may be due to (a) setting multiple cutoffs, (b) the military testing situation, (c) setting standards at the test level, or (d) some combination of the three. Additional research is needed to determine whether these factors or others cause an increase in standards.

Examinee-Based Archival Technique

Cantor (1989) demonstrated that examinee-based procedures could be used to set standards on archival data (i.e., without the use of SME judgments). Given budget constraints and time resources necessary to conduct standard setting workshops, a similar use of archival data seems practical for establishing performance standards in the present project. Specifically, rating data could be used to map cutoffs on the hands-on tests.

The examinee-based archival technique takes advantage of the existing JPM database. The database contains hands-on test score (HOTS) information and job performance ratings for 261 AGE personnel. Two types of job performance ratings were collected: Global Technical Proficiency (GTP) and Technical Proficiency (TP). GTP ratings refer to how skilled an individual is at performing the technical aspects of the job ignoring interpersonal factors (i.e., willingness to work, cooperation with others) or situational factors (i.e., lack of tools, parts, or equipment). The GTP rating is an evaluation of the quality of an individual's work across tasks. On the other hand, TP ratings refer to how skilled an individual is at performing a specific task. By definition, TP ratings exclude interpersonal and situational factors. TP ratings for each of the tasks are based on the question, "At what level of proficiency could this individual perform this particular task?" Thus, the GTP rating is an overall rating of technical proficiency (i.e., across all technical aspects of the job), whereas the TP rating points to proficiency on one task in particular.

TP ratings for each of the 16 tasks and GTP ratings were made using the 5-point rating scale presented in Figure 1. GTP and TP ratings were obtained from four rating sources: (a) test administrator (TA), (b) incumbent's supervisor (S), (c) examinee or incumbent (I), and (d) up to three of the incumbent's coworkers or peers (P_{1-3}). Note that there are four rating sources (TA, S, I, and P) but the number of raters for any one incumbent may be as low as four (TA, S, I, P_1) or as high as six (TA, S, I, P_1, P_2, P_3). GTP ratings were available from the incumbent, supervisor, and peers, but not from the test administrator. By nature, GTP ratings summarize performance observed across time and tasks. The test administrator does not interact daily with the examinee; therefore, he or she is not qualified to provide GTP ratings. TP ratings were available from all four sources. The test administrator's TP ratings were recorded immediately after observing the examinee perform the hands-on test. The incumbent, supervisor, and peers, on the other hand, relied on recall of past performance when providing TP ratings.

Originally, we proposed combining the two types of ratings (GTP, TP) obtained from the four sources (TA, S, I, and P) in four different ways (Fotouhi et al., 1990). However, earlier work by Hedge and Ringenbach (1989) demonstrated that the GTP ratings were unreliable due to halo error and range restriction. Thus, upon consultation with AFHRL representatives, the GTP ratings were dropped from further examination.

In using the TP ratings to establish standards on the hands-on tests, a mean hands-on test score (MHOTS) had to first be calculated by rater (TA, S, I, P_1, P_2, P_3) and proficiency level (Unqualified, Marginal, Qualified, Distinguished, and Exceptional). Figure 5 graphically presents an example of how one incumbent's (I_A) test score was used in defining the proficiency level for a single hands-on test (T_1). Suppose for T_1 I_A was rated "Qualified" by the TA, "Distinguished" by the S and I, "Qualified" by P_1 and P_2, and "Distinguished" by P_3. The HOTS of I_A would contribute to the calculation of

Source Proficiency Levels

	Unqualified	Marginal	Qualified	Distinguished	Exceptional
TA			I_A		
S				I_A	
I				I_A	
P_1			I_A		
P_2			I_A		
P_3				I_A	

Figure 5. Calculation of mean hands-on test scores (MHOTSs) by rater and proficiency level.

six MHOTSs--an MHOTS corresponding to each rater and the rater's designated proficiency level (TA-"Qualified", S-"Distinguished", I-"Distinguished", P_1-"Qualified", P_2-"Qualified", and P_3-"Distinguished"). A total of 30 (six raters by five proficiency levels) MHOTSs for each of the 16 tests were calculated in this manner.

The MHOTSs by proficiency level were combined in three ways as shown in Table 6. The first, labelled TP-TA, is the TA derived MHOTS for each proficiency level. The second, labelled TP-SPI, is an average MHOTS calculated from the S, P, and I MHOTSs for each proficiency level. (Recall that as many as three peer ratings were available for any one incumbent. An incumbent's peer MHOTSs were averaged to yield a single peer MHOTS for each proficiency level.) The final combination, labelled TP-ALL, is the MHOTS across all sources for each proficiency level.

Method for Calculating Cutoffs

To define the boundaries between Unqualified, Marginal, Qualified, Distinguished, and Exceptional performance, the mid-point between the MHOTSs for each proficiency level was determined. For example, consider Test 154--Perform an Aircraft Support Generator Service Inspection and the TP-TA categorization method. The MHOTS for individuals categorized as Unqualified is 5.123, for Marginal is 7.333, for Qualified is 8.196, for Distinguished is 9.274, and for Exceptional is 9.936. The mid-point score between the mean Unqualified (5.123) and mean Marginal (7.333) scores is 6.23. Thus, 6.23 is the lower boundary for Marginal performance. In other words, to be classified as Marginal on Test 154, an airman must receive a HOTS between 6.23 and 7.77 (the midpoint between Marginal and Qualified). Any score below 6.23 defines Unqualified performance. The lower boundaries (i.e., cutoffs) for Qualified, Distinguished, and Exceptional performance are 7.77, 8.74, and 9.61, respectively, for Test 154.

Table 6

Methods of Categorizing Examinees into Five Proficiency Levels and Calculating Mean Hands-on Test Score (MHOTS) at Each Level

Label	Source(s)	Mean Hands-on Test Score for the Five Proficiency Levels
TP-TA	TA	MHOTS at each proficiency level
TP-SPI	S, P, I	MHOTS at each proficiency level averaged across sources
TP-ALL	TA, S, P, I	MHOTS at each proficiency level averaged across sources

Note. TP = Technical Proficiency; TA = Test Administrator; S = Supervisor; P = Peer; I = Incumbent.

Correction for "missing" data. As in any large scale data collection, or in dealing with a database, there seems to be missing data. In this case, "missing" is defined as a lack of ratings by a particular rating source (TA, S, P, or I) at a specific proficiency level (i.e., a truncated range). In other words, the type of rater simply did not rate anyone at that particular level. In one case of "missing" data, the TAs did not rate any incumbents at the Exceptional proficiency level (i.e., Test 179--Perform a Gas Turbine Compressor Periodic Inspection). In all other instances of "missing" data, at least one rater (TA, S, I, P_1, P_2, or P_3) did not rate any incumbents at the Unqualified proficiency level. The "missing" data at the Unqualified rating level did not affect the calculation of the boundaries between the Unqualified and Marginal levels of performance. To assist in graphing the data for the tests where there was "missing" data at the Unqualified level, an average MHOTS from the other raters was substituted for the "missing" data. Thus, the average MHOTS for incumbents rated Unqualified did not change by this procedure. Table 7 displays the procedure by which the value for the missing data was adjusted and the lack of change in both the MHOTSs for each proficiency level and the boundaries between each proficiency level.

In the case where the TAs did not rate anyone as Exceptional (Test 179--Perform a Gas Turbine Compressor Periodic Inspection), no boundary could be established between the Distinguished and the Exceptional levels using the TP-TA method or the TP-ALL method without a MHOTS for the Exceptional level. An average MHOTS across the rating sources for the Exceptional level could not be substituted for the "missing" data because the Distinguished level was higher than the Exceptional level. The cutoff for the lower level (i.e., Distinguished) cannot be greater than the next higher level (i.e., Exceptional). In this case, the data were "adjusted" so that the TA derived MHOTS for the Distinguished level of performance was used at both the Distinguished and Exceptional levels. The Exceptional cutoff set by the TP-TA method was

Table 7

Example of Adjustment for Missing Data for Test 154--Perform an Aircraft Support Generator Service Inspection

	Proficiency Level Mean Hands-on Test Score				
Source	Unqualified	Marginal	Qualified	Distinguished	Exceptional
TA	5.1232	7.3334	8.1966	9.2740	9.9362
S	6.9666*	7.7506	7.9356	8.3712	7.9735
I	8.0000	6.9340	7.8136	8.1461	8.5284
P_{1-3}	7.7767	8.0770	7.8864	8.1171	8.2511
MHOTS Across Rating Sources w/o Adj.	6.9666	7.5238	7.9581	8.4771	8.6723
Cutoffs w/o Adj.		7.2452	7.7409	8.2176	8.5747
MHOTS Across Rating Sources with Adj.	6.9666	7.5238	7.9581	8.4771	8.6723
Cutoffs with Adj.		7.2452	7.7409	8.2176	8.5747

* Adjusted MHOTS.

"adjusted" so that anyone scoring above the MHOTS at the Distinguished level would be considered Exceptional. For Test 179 the MHOTS for incumbents who were rated Distinguished is 6.62. It can be assumed that incumbents who were rated Exceptional would score at or above 6.62. Thus, this "adjustment" produces increasing cutoffs, which are consistent with the pattern of increasing levels of proficiency. Table 8 displays the "adjustment" method and the increasing step levels of the cutoffs.

Results and Discussion

Table 9 presents cutoffs for the 16 hands-on tests by performance level and classification procedure. The mean cutoff across all tests is presented at the bottom of Table 9. TP ratings were not provided by supervisors, peers, or incumbents for Test 284--Remove and Replace Engine Fan Belts; therefore, the TP-SPI and TP-ALL methods could not be used to determine cutoffs for that test. Figure 6 presents a graphic comparison of the standards set across tests using the three classification procedures.

Differences among methods. An examination of the data in Table 9 and Figure 6 indicates that the classification methods do not produce the same standards. Furthermore, the TP-SPI and TP-ALL classification methods appear to yield cutoffs (Marginal, Qualified, Distinguished, and Exceptional) that are not significantly different. Repeated measures analyses of variance (ANOVAs) were used to test differences between classification methods and

Table 8

Example of Adjustment for Missing Data for Test 179--Perform a Gas Turbine Compressor Periodic Inspection

	Proficiency Level Mean Hands-on Test Score								
Source	Unqualified		Marginal		Qualified		Distinguished		Exceptional
TA	2.3730		4.6481		6.5063		6.6200		6.6200*
S	3.3767		2.4554		2.9728		3.3193		3.3089
I	2.4256		2.1569		2.7566		3.3683		3.8843
P_{1-3}	1.1700		2.3095		2.9855		3.5460		3.5862
MHOTS Across Rating Sources w/o Adj.	2.3363		2.8925		3.8053		4.2134		3.5931
Cutoffs w/o Adj.		2.6144		3.3489		4.0093		3.9033	
MHOTS Across Rating Sources with Adj.	2.3363		2.8925		3.8053		4.2134		4.3499
Cutoffs with Adj.		2.6144		3.3489		4.0093		4.2816	

* Adjusted MHOTS.

cutoffs. Analyses were conducted with the hands-on tests as cases and the cutoffs as the repeated "trials." TP-TA cutoffs for Test 284--Remove and Replace Engine Fan Belts were omitted from the ANOVAs because the TP-SPI and TP-ALL classification methods were not used to set standards for that test. Significant main effects were found for method ($\underline{F}_{(2,28)}$=43.431, $\underline{p}$=.000) and cutoff ($\underline{F}_{(3,42)}$=195.735, $\underline{p}$=.000), and the interaction was also significant ($\underline{F}_{(6,84)}$=86.066, $\underline{p}$=.000).

Based on the ANOVA results, planned comparisons were made to examine differences among the classification methods. Table 10 presents the results of planned comparisons for the TP-TA, TP-SPI, and TP-ALL classification methods. Results indicate that the three methods produce significantly different standards.

Differences among raters. The next step was to identify the most appropriate set of standards: Those produced by the TP-TA method, the TP-SPI method, or the TP-ALL method. In order to determine the most appropriate standards, it was first necessary to identify the ratings which were the most accurate in predicting job proficiency. We received an indication early on that the TA ratings were more accurate than the S, I, and P ratings based on the generation of graphs. An example of these graphs is in Figure 7.

The graph in Figure 7 shows the MHOTS across the five proficiency levels for the different rating sources (i.e., TA, S, I, P) and the combination of

Table 9

Archival Technique: Minimum Test Score Cutoff by Test and Classification Method

	Marginal			Qualified			Distinguished			Exceptional		
Test	TP-TA	TP-SPI	TP-ALL	TP-TA	TP-SPI	TP-ALL	TP-TA	TP-SPI	TP-ALL	TP-TA	TP-SPI	TP-ALL
154	6.23	7.54	7.21	7.77	7.73	7.74	8.74	8.05	8.22	9.61	8.23	8.57
155	4.48	4.88	4.78	6.32	4.61	5.04	7.93	4.62	5.44	9.03	4.64	5.74
162	6.08	6.34	6.27	7.61	7.14	7.26	8.50	7.39	7.67	9.35	7.67	8.09
179	3.51	2.32	2.61	5.58	2.61	3.35	6.56	3.16	4.01	6.62	3.50	4.28
209	5.19	5.67	5.55	7.17	6.01	6.30	8.18	6.59	6.98	9.17	6.95	7.50
215	4.75	4.56	4.61	7.79	4.94	5.65	8.85	6.20	6.86	9.69	6.96	7.64
238	6.22	6.68	6.57	7.99	6.90	7.17	9.28	7.28	7.78	9.78	7.49	8.07
251	3.57	1.69	2.16	7.21	2.41	3.61	8.45	3.08	4.42	9.33	4.07	5.38
260	4.42	4.80	4.70	7.88	5.65	6.21	8.84	6.37	6.99	9.55	6.69	7.40
264	4.81	4.16	4.33	6.23	4.89	5.23	7.84	5.33	5.96	9.40	5.35	6.36
284	5.74	--	5.74	7.69	--	7.69	8.89	--	8.89	9.76	--	9.76
300	6.19	7.17	6.92	7.61	7.33	7.40	8.73	7.50	7.81	9.57	7.64	8.12
421	2.89	3.03	3.00	4.68	2.87	3.32	6.71	3.13	4.03	8.84	3.18	4.60
446	3.28	2.43	2.64	5.61	2.63	3.37	7.17	2.84	3.92	8.90	3.10	4.55
503	4.66	5.19	5.06	6.36	5.32	5.58	8.03	5.44	6.09	9.44	5.64	6.59
549	4.94	5.19	5.12	6.79	6.23	6.37	8.26	6.34	6.82	9.38	6.66	7.34
Across all Tests	4.81	4.78	4.83	6.89	5.15	5.71	8.19	5.55	6.37	9.21	5.85	6.87

Note. TP = Technical Proficiency; TA = Test Administrator; S = Supervisor; P = Peer; I = Incumbent; -- = No Ratings for S, P, or I therefore TP-ALL = TP-TA.

Tests are:

154 = Perform Aircraft Support Generator Service Insp.
155 = Service Inspection on a Load Bank
162 = Perform Service Inspection on Hydraulic Test Stand
179 = Perform Gas Turbine Compressor Periodic Inspection
209 = Measure Resistance in AGE Electrical Systems
215 = Perform AGE Electrical Systems Operational Checks
238 = Splice Electrical Systems Wiring
251 = Adjust Turbine Engine Fuel System Components
260 = Clean Motor and Generator Components
264 = Isolate Engine, Motor, or Generator Malf.
284 = Remove and Replace Engine Fan Belts
300 = Remove or Install Fuel Lines and Fittings
421 = Remove or Install Hydraulic Lines
446 = Isolate Pneumatic System Malfunctions
503 = Research T.O.s for AGE Chassis
549 = Inspect Vehicles for Safety of Operation

Table 10

Planned Comparisons for Method Classification Using Repeated Measures ANOVA

Source	SS	df	MS	F	p
TP-TA vs. TP-SPI	859.422	1	859.422	43.423	.000
Error	277.088	14	19.792		
TP-TA vs. TP-ALL	483.822	1	483.822	43.459	.000
Error	155.861	14	11.133		
TP-SPI vs. TP-ALL	53.582	1	53.582	43.311	.000
Error	17.320	14	1.237		

Note. TP = Technical Proficiency; TA = Test Administrator; SPI = Supervisor, Peer, and Incumbent; ALL = Test Administrator, Supervisor, Peer, and Incumbent.

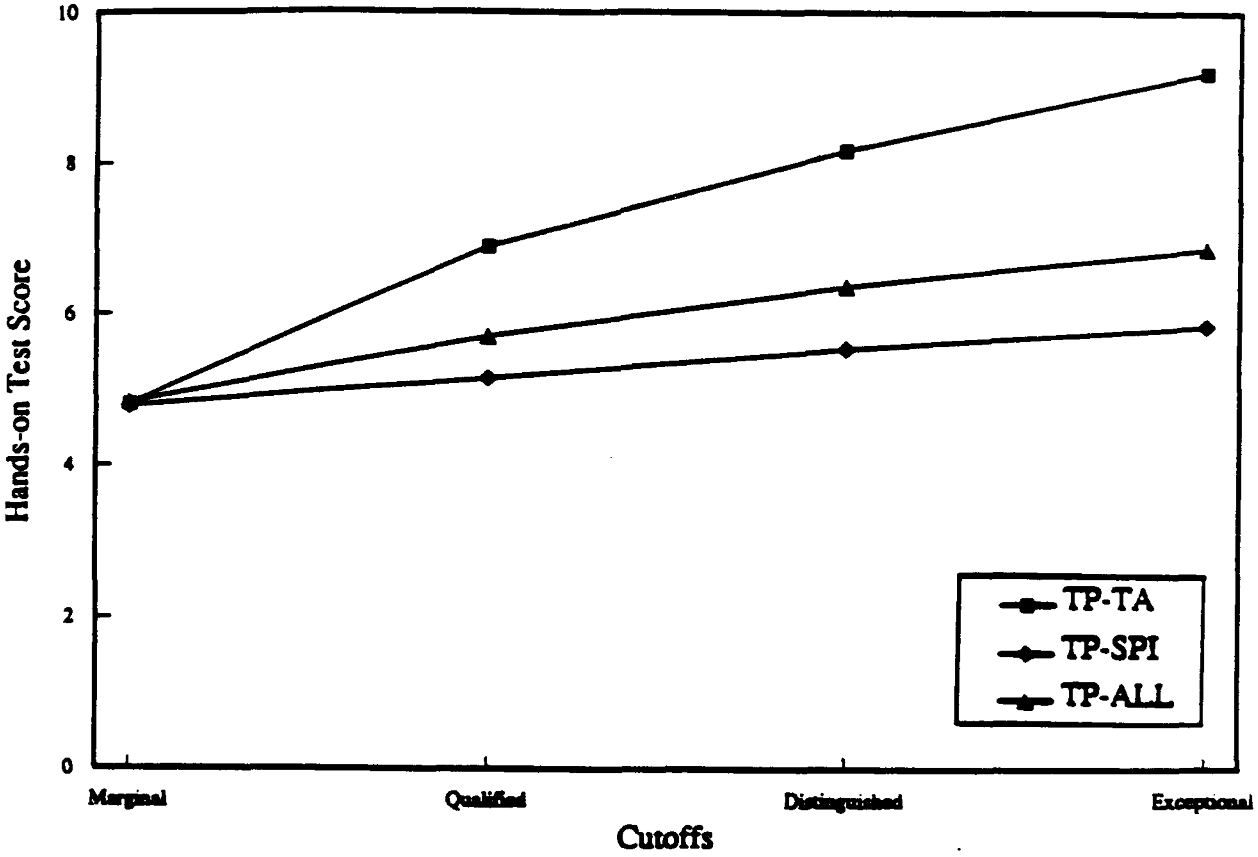

Figure 6. Archival technique: Standards across all hands-on tests set via three classification procedures.

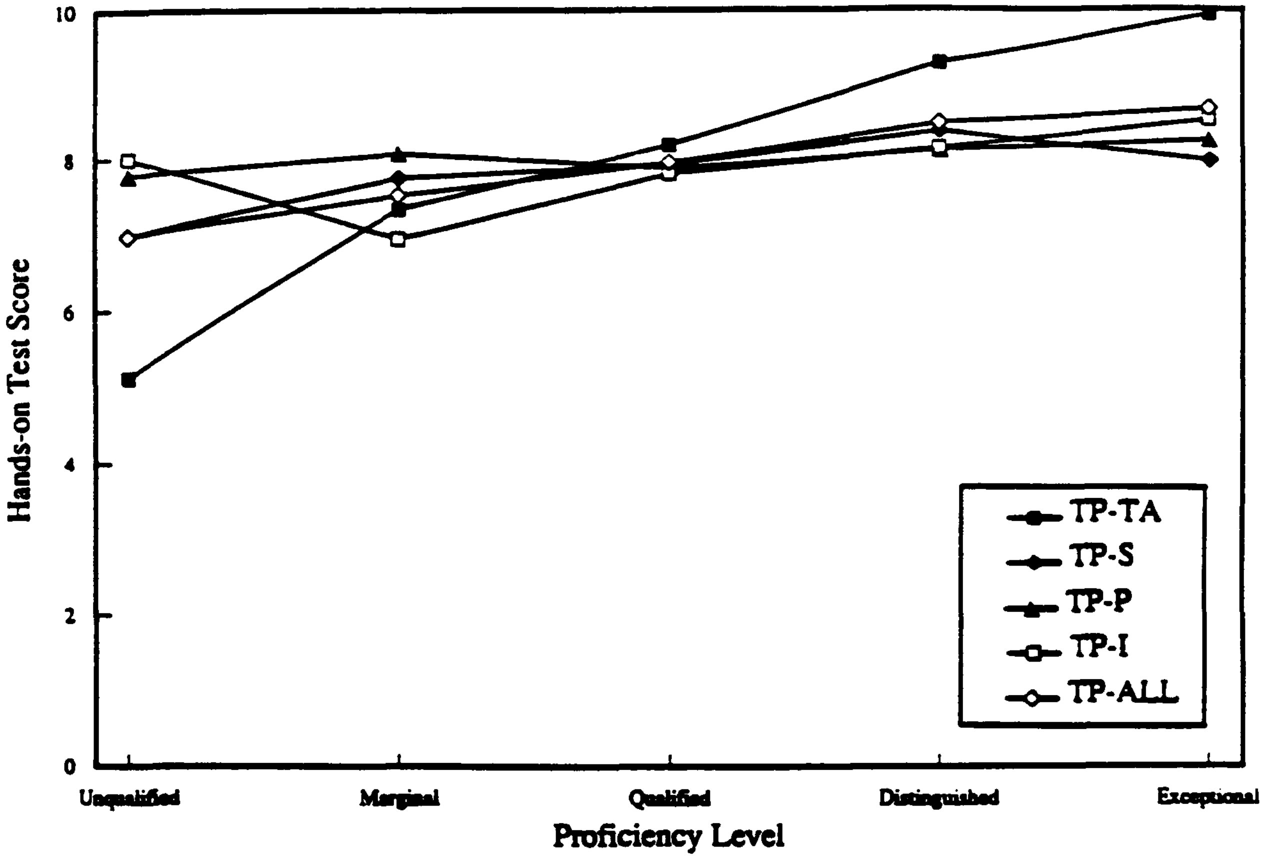

Figure 7. Archival technique: Technical Proficiency (TP) level by hands-on test score via separate sources for Test 154--Perform an Aircraft Support Generator Service Inspection.

all sources (i.e., ALL). Consistently, the slope of the TA line across the proficiency levels is greater than that of the other raters. In other words, TAs rated the poor performers (low HOTS) low on the proficiency scale and the good performers (high HOTS) high on the proficiency scale. The slope is an indication of the rater's ability to separate the poor from the good performers. The greater the slope, the greater the variability in the HOTSs between the five performance levels.

Pearson correlation coefficients between TA, S, P, and I ratings and HOTSs were then computed for each of the 16 hands-on tests. As can be seen in Table 11, in every case the TA correlated significantly with the HOTS with the lowest correlation coefficient being .54 (p=.000) and the highest being .86 (p=.000). The other raters did not correlate as highly with the HOTS; the highest correlation coefficient being .34 for the supervisors on Test 215--Perform AGE Electrical Systems Operational Checks. While some of the correlations of the S, I, and P ratings with HOTS were significant, the correlations were small in comparison to the TA correlations. In light of the high correlations of the TA ratings, other analyses of the rating data were

Table 11

Archival Technique: Correlation of Test Administrator (TA), Supervisor (Super), Incumbent (Incum), and Peer Ratings with Hands-on Test Score

Test		Correlation with Hands-on Test Score					
		TA	Super	Incum	Peer 1	Peer 2	Peer 3
154	Coef.	.74	.07	.20	.10	.00	.03
	1 Tail Sig.	.000	.200	.008	.101	.488	.376
155	Coef.	.71	.02	.05	.17	.01	-.09
	1 Tail Sig.	.000	.427	.291	.026	.472	.155
162	Coef.	.75	.14	.12	.07	.11	.13
	1 Tail Sig.	.000	.044	.073	.205	.094	.050
179	Coef.	.54	.10	.26	.31	.31	.08
	1 Tail Sig.	.000	.111	.001	.000	.000	.159
209	Coef.	.68	.24	.22	.30	.28	.14
	1 Tail Sig.	.000	.001	.004	.000	.000	.049
215	Coef.	.77	.34	.27	.12	.13	.16
	1 Tail Sig.	.000	.000	.000	.066	.049	.022
238	Coef.	.67	.05	.14	.06	.12	.06
	1 Tail Sig.	.000	.283	.047	.234	.070	.229
251	Coef.	.86	.19	.23	.25	.27	.11
	1 Tail Sig.	.000	.010	.002	.001	.000	.095
260	Coef.	.84	.01	.25	.21	.19	.07
	1 Tail Sig.	.000	.460	.001	.004	.009	.188
264	Coef.	.65	.22	.13	.08	.12	.07
	1 Tail Sig.	.000	.004	.063	.176	.065	.182
284	Coef.	.78	---	---	---	---	---
	1 Tail Sig.	.000					
300	Coef.	.77	.11	.06	.03	.08	.08
	1 Tail Sig.	.000	.086	.245	.370	.152	.167
421	Coef.	.69	.03	.12	.10	.11	-.09
	1 Tail Sig.	.000	.366	.071	.102	.097	.146
446	Coef.	.81	.02	.10	.11	.16	-.04
	1 Tail Sig.	.000	.414	.108	.095	.028	.322

Table 11 (continued)

Test		Correlation with Hands-on Test Score					
		TA	Super	Incum	Peer 1	Peer 2	Peer 3
503	Coef.	.81	.05	.16	-.01	.01	.03
	1 Tail Sig.	.000	.255	.024	.478	.465	.352
549	Coef.	.82	.09	.21	.16	.11	.11
	1 Tail Sig.	.000	.125	.004	.026	.087	.086

Note. --- = No Ratings Supplied.

performed.

Single factor ANOVAs were conducted for each hands-on test to examine the differences between the five performance levels for each rating source. The dependent variable was defined as the MHOTS, and the independent variable was defined as the particular rater. With these ANOVAs we hoped to find significant differences between MHOTS at each performance level. If the MHOTS was not significantly different across rating levels, which the graphs like Figure 7 led us to believe, then that particular rater did not accurately place incumbents into the proficiency level suggested by their HOTS. For every test, the difference between the rating levels of the TA were significant at the p=.000 level. However, the results for the other raters were mixed. While significant results were found for some of the raters, no clear pattern of significance was found for the S, I, or P ratings across tasks. Our purpose for running these analyses was to recommend either a certain rater or combination of raters to be used to classify the incumbents into the five levels of proficiency. Then the MHOTSs for the proficiency levels could be calculated. Midpoints between those MHOTSs could be generated to set the standards at each of the five levels of proficiency.

Separately, only the TA consistently provided ratings which produced variability in the MHOTSs between the five proficiency levels. To determine if a combination of raters would also produce the variability needed to recognize five distinct levels of proficiency, we conducted a multiple factor ANOVA for each test with the raters as the independent variables. The dependent variable was again the MHOTS. The main effect of the raters again showed the TA producing significant differences at the p=.000 level. Table 12 summarizes the results.

The multiple factor ANOVAs required that a proficiency level rating be provided by each rating source (TA, S, I, P_1, P_2, P_3) for the incumbents. This reduced the sample size from 261 incumbents to approximately 150 across tasks. The TA, again, consistently accounted for significant variation in test performance (p=.000). There were only three cases in which any other rater accounted for significant variation in HOTS: (a) Test 251, S at p=.005,

Table 12

Results of Multiple Factor ANOVAs of Hands-on Test Score by Rater

Test	Main Effects	SS	df	MS	*F*	*p*
154	TA	217.007	4	54.252	48.765	.000
	S	.752	3	.251	.225	.879
	I	3.881	4	.970	.872	.483
	P_1	3.937	4	.984	.885	.475
	P_2	4.749	4	1.187	1.067	.376
	P_3	2.058	4	.515	.463	.763
	Explained	260.518	23	11.327	10.181	.000
	Residual	141.290	127	1.113		
155	TA	181.781	4	45.445	28.633	.000
	S	10.018	4	2.504	1.578	.186
	I	5.946	4	1.486	.937	.446
	P_1	4.277	4	1.069	.674	.612
	P_2	3.477	4	.869	.548	.701
	P_3	10.118	4	2.529	1.594	.181
	Explained	266.650	24	11.110	7.000	.000
	Residual	168.240	106	1.587		
162	TA	191.017	4	47.754	38.068	.000
	S	10.986	4	2.746	2.189	.074
	I	3.505	4	.876	.698	.594
	P_1	1.372	4	.343	.273	.895
	P_2	4.458	4	1.114	.888	.473
	P_3	3.525	3	1.175	.937	.425
	Explained	271.726	23	11.814	9.418	.000
	Residual	159.316	127	1.254		
179	TA	108.661	3	36.220	15.393	.000
	S	1.694	4	.424	.180	.948
	I	4.535	4	1.134	.482	.749
	P_1	3.704	4	.926	.394	.813
	P_2	17.492	4	4.373	1.858	.122
	P_3	18.357	4	4.589	1.950	.106
	Explained	228.702	23	9.944	4.226	.000
	Residual	301.187	128	2.353		

Table 12 (continued)

Test	Main Effects	SS	df	MS	*F*	*p*
209	TA	249.763	4	62.441	33.266	.000
	S	9.835	4	2.459	1.310	.270
	I	9.971	4	2.493	1.328	.263
	P_1	7.875	4	1.969	1.049	.385
	P_2	8.622	4	2.156	1.148	.337
	P_3	14.565	4	3.641	1.940	.108
	Explained	382.521	24	15.938	8.491	.000
	Residual	234.624	125	1.877		
215	TA	1015.269	4	253.817	66.697	.000
	S	10.199	4	2.550	.670	.614
	I	7.876	4	1.969	.517	.723
	P_1	6.470	4	1.618	.425	.790
	P_2	1.654	3	.551	.145	.933
	P_3	10.256	4	2.564	.674	.611
	Explained	1464.458	23	63.672	16.731	.000
	Residual	487.109	128	3.806		
238	TA	282.899	4	70.725	28.512	.000
	S	7.021	4	1.755	.708	.588
	I	10.391	4	2.598	1.047	.386
	P_1	3.223	3	.741	.299	.826
	P_2	11.629	3	3.876	1.563	.202
	P_3	4.683	3	1.561	.629	.596
	Explained	359.429	21	17.116	6.900	.000
	Residual	319.993	129	2.481		
251	TA	1321.127	4	330.282	358.871	.000
	S	14.633	4	3.658	3.975	.005
	I	7.204	4	1.801	1.957	.105
	P_1	.691	4	.173	.188	.944
	P_2	6.155	4	1.539	1.672	.161
	P_3	10.294	4	2.731	2.968	.022
	Explained	1666.527	24	69.439	75.449	.000
	Residual	115.962	126	.920		

Table 12 (continued)

Test	Main Effects	SS	df	MS	F	p
260	TA	1509.531	4	377.383	250.010	.000
	S	.995	4	.249	.165	.956
	I	6.803	4	1.701	1.127	.347
	P_1	14.961	4	3.740	2.478	.047
	P_2	6.149	4	1.537	1.018	.400
	P_3	11.794	4	2.948	1.953	.106
	Explained	1887.675	24	78.653	52.106	.000
	Residual	191.703	127	1.509		
264	TA	123.768	3	41.256	32.535	.000
	S	5.668	4	1.417	1.117	.351
	I	18.361	4	4.590	3.620	.008
	P_1	11.367	4	2.842	2.241	.068
	P_2	4.869	4	1.217	.960	.432
	P_3	9.045	4	2.261	1.783	.136
	Explained	193.180	23	8.399	6.624	.000
	Residual	161.042	127	1.268		
284	TA	537.016	4	134.254	131.289	.000
	Explained	537.016	4	134.254	131.289	.000
	Residual	261.782	256	1.023		
300	TA	209.801	4	52.450	53.710	.000
	S	1.019	3	.340	.348	.791
	I	2.966	3	.989	1.012	.390
	P_1	2.158	3	.719	.736	.532
	P_2	4.018	4	1.005	1.029	.395
	P_3	1.313	4	.328	.336	.853
	Explained	242.041	21	11.526	11.803	.000
	Residual	124.997	128	.977		

Table 12 (continued)

Test	Main Effects	SS	df	MS	*F*	*p*
421	TA	210.264	3	70.088	36.078	.000
	S	1.371	3	.457	.235	.872
	I	2.380	4	.595	.306	.873
	P_1	6.867	4	1.717	.884	.476
	P_2	6.035	4	1.509	.777	.542
	P_3	9.758	3	3.253	1.674	.176
	Explained	290.207	21	13.819	7.114	.000
	Residual	250.604	129	1.943		
446	TA	515.531	4	128.883	64.508	.000
	S	1.349	4	.337	.169	.954
	I	11.529	4	2.882	1.443	.224
	P_1	20.167	4	5.042	2.524	.044
	P_2	13.832	4	3.458	1.731	.147
	P_3	2.919	4	.730	.365	.833
	Explained	661.655	24	27.569	13.799	.000
	Residual	247.745	124	1.998		
503	TA	376.477	4	94.119	61.497	.000
	S	3.422	4	.856	.559	.693
	I	1.036	4	.259	.169	.954
	P_1	10.773	4	2.693	1.760	.141
	P_2	.874	4	.218	.143	.966
	P_3	2.321	4	.580	.379	.823
	Explained	417.162	24	17.382	11.357	.000
	Residual	194.368	127	1.530		

Note. Test 284 contained only TA ratings. Test 549 analyses could not be completed due to singular matrix for main effects and covariates.
Tests are:
154 = Perform Aircraft Support Generator Service Inspection
264 = Isolate Engine, Motor, or Generator Malfunction
155 = Perform Service Inspection on a Load Bank
284 = Remove and Replace Engine Fan Belts
162 = Perform Service Inspection on Hydraulic Test Stand
300 = Remove or Install Fuel Line and Fittings
179 = Perform Gas Turbine Compressor Periodic Inspection
421 = Remove or Install Hydraulic Lines
209 = Measure Resistance in AGE Electrical Systems
446 = Isolate Pneumatic System Malfunctions
215 = Perform AGE Electrical Systems Operational Checks
503 = Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance

Table 12 (continued)

Information
238 = Splice Electrical Systems Wiring
251 = Adjust Turbine Engine Fuel System Components
549 = Inspect Vehicles for Safety of Operation
260 = Clean Motor and Generator Components

(b) Test 251, P_3 at p=.022, and (c) Test 264, I at p=.008. The overall p value for each test, however, gives significant results at the p=.000 level. So there is variability across the five levels of proficiency when all the raters are included. It was therefore necessary to determine which of the raters were accounting for the variance found to be significant in the ANOVAs.

Multiple regression analyses were then conducted to specify the rater or raters who were accounting for the largest amount of variance and providing the best predictions of job proficiency. We strongly believed that the TA was accounting for most of the variance. The previously run ANOVAs showing significant results for all TAs, non-significant results for all but a handful of the other raters, and significant overall results for every task made us believe that all the raters combined provided no more information than the information provided by the TA alone. With this in mind, the regression models specified that the TA ratings were entered first and run against the MHOTS for each task. The other raters were then entered in the model to test for the change in the variance accounted for. Table 13 summarizes the results.

What we are seeking to show with these regression analyses is not a mere description of the data; rather we wish to make decisions based on these results. To do that, however, we must first describe the specific values resulting from the regression. R^2 is known as the coefficient of determination. Strictly speaking, it is the square of the correlation coefficient between a criterion and some number of predictor variables. The adjusted R^2 takes into consideration the number of predictors in the model. Thus, this value is an approximation of the R^2 one would expect if this model were fit to the population from which it was sampled.

In our case, we have two models to compare to the population. The first is the regression of the TA ratings in predicting the HOTS of the incumbents. The second is a model that includes all the raters. Remember that we already had an idea that the TA ratings accounted for the most variation in hands-on test performance. Comparison of these two models allows an assessment of how much the other ratings contribute to the ratings made by the TA for predicting job proficiency. To illustrate the use of these statistics, we will use Test 154--Perform Aircraft Support Generator Service Inspection as an example. The R^2 change statistic is the increase in R^2 observed when the rest of the ratings are added to the TA model. Test 154 shows an increase in R^2 of .0087. The F change is a test of the increase in R^2. In Test 154, this increase is not significant. The T statistic tests the linear relationship between the variables. It points, in our case, to the

Table 13

Results of the Linear Regression of Mean Hands-on Test Score and the Individual Raters

Test		TA	ALL	Source	Beta	SE Beta	T	Sig T
154	R^2	.5525	.5612	TA	1.1589	.0894	12.959	.0000
	Adj. R^2	.5495	.5429	S	-.0615	.1264	-.486	.6275
				I	.1037	.1041	.996	.3207
	Effect of Adding Raters			P_1	-.0508	.1168	-.435	.6645
	R^2 Change =		.0087	P_2	.1102	.1235	.892	.3739
	F Change =		.5695	P_3	-.0988	.1162	-.850	.3967
	F Change Sig. =		.7233					
155	R^2	.5092	.5299	TA	1.5653	.0894	11.185	.0000
	Adj. R^2	.5054	.5072	S	-.1138	.1258	-.905	.3674
				I	.1029	.1155	.891	.3747
	Effect of Adding Raters			P_1	.1860	.1334	1.395	.1657
	R^2 Change =		.0207	P_2	.0278	.1459	.191	.8491
	F Change =		1.0910	P_3	-.2070	.1231	-1.682	.0951
	F Change Sig. =		.3688					
162	R^2	.5573	.5671	TA	1.1199	.0846	13.239	.0000
	Adj. R^2	.5543	.5491	S	.0767	.1076	.713	.4772
				I	-.0369	.1193	-.309	.7575
	Effect of Adding Raters			P_1	-.0537	.1172	-.458	.6478
	R^2 Change =		.0098	P_2	.1352	.1223	1.106	.2706
	F Change =		.6510	P_3	.0947	.1231	.769	.4432
	F Change Sig. =		.6612					
179	R^2	.2925	.3628	TA	1.7357	.2601	6.673	.0000
	Adj. R^2	.2878	.3364	S	-.0357	.1486	-.240	.8105
				I	.1154	.1387	.832	.4066
	Effect of Adding Raters			P_1	.2053	.1503	1.366	.1742
	R^2 Change =		.0703	P_2	.4362	.1681	2.594	.0104
	F Change =		3.1988	P_3	.0397	.1499	.265	.7915
	F Change Sig. =		.0091					
209	R^2	.4670	.5010	TA	1.3713	.1375	9.973	.0000
	Adj. R^2	.4634	.4800	S	.1021	.1345	.760	.4488
				I	.1180	.1373	.859	.3915
	Effect of Adding Raters			P_1	.2070	.1371	1.510	.1332
	R^2 Change =		.0340	P_2	.2098	.1587	1.322	.1884
	F Change =		1.9474	P_3	-.0517	.1483	-.349	.7276
	F Change Sig. =		.0901					

Table 13 (continued)

Test		TA	ALL	Source	Beta	SE Beta	T	Sig T
215	R^2	.5887	.6141	TA	2.1440	.1636	13.104	.0000
	Adj. R^2	.5859	.5981	S	.5321	.2192	2.428	.0164
				I	.3007	.2305	1.304	.1942
	Effect of Adding Raters			P_1	-.0423	.2004	-.211	.8331
	R^2 Change =		.0254	P_2	.0430	.2309	.186	.8526
	F Change =		1.9081	P_3	-.0515	.2369	-.217	.8283
	F Change Sig. =		.0965					
238	R^2	.4427	.4488	TA	1.2242	.1170	10.467	.0000
	Adj. R^2	.4389	.4258	S	-.0443	.1523	-.291	.7716
				I	.1184	1522	.778	.4379
	Effect of Adding Raters			P_1	.0376	.1582	.238	.8125
	R^2 Change =		.0061	P_2	.1343	.1688	.795	.4277
	F Change =		.3201	P_3	-.0658	.1624	-.405	.6858
	F Change Sig. =		.9002					
251	R^2	.7408	.7643	TA	3.0131	.1528	19.726	.0000
	Adj. R^2	.7391	.7545	S	.1258	.1719	.732	.4657
				I	.2737	.1706	1.604	.1109
	Effect of Adding Raters			P_1	.1681	.2022	.832	.4070
	R^2 Change =		.0235	P_2	.3207	.1928	1.664	.0948
	F Change =		2.8654	P_3	.1187	.1641	.723	.4707
	F Change Sig. =		.0169					
260	R^2	.7104	.7211	TA	2.3023	.1287	17.887	.0000
	Adj. R^2	.7084	.7095	S	-.2512	.1606	-1.564	.1200
				I	.1766	.1738	1.106	.3115
	Effect of Adding Raters			P_1	-.0151	.1848	-.082	.9349
	R^2 Change =		.0107	P_2	-.0431	.1982	-.217	.8283
	F Change =		1.1145	P_3	-.2218	.1891	-1.173	.2429
	F Change Sig. =		.3553					
264	R^2	.4253	.4421	TA	1.4098	.1412	9.984	.0000
	Adj. R^2	.4214	.4189	S	.1138	.1069	1.065	.2886
				I	.0785	.1288	.609	.5433
	Effect of Adding Raters			P_1	-.1150	.1150	-1.000	.3197
	R^2 Change =		.0168	P_2	.1303	.1249	1.043	.2987
	F Change- =		.8690	P_3	.0237	.1166	.203	.8391
	F Change Sig. =		.5036					

Table 13 (continued)

Test		TA	ALL	Source	Beta	SE Beta	T	Sig T
300	R^2	.5996	.6021	TA	1.8327	.0819	14.452	.0000
	Adj. R^2	.5969	.5854	S	.0553	.1087	.508	.6120
				I	-.0397	.1090	-.364	.7165
	Effect of Adding Raters			P_1	-.0103	.1103	-.093	.9259
	R^2 Change =		.0025	P_2	.0590	.1103	.535	.5932
	F Change =		.1796	P_3	-.0566	.1036	-.547	.5855
	F Change Sig. =		.9699					
421	R^2	.4732	.4926	TA	1.8192	.1615	11.267	.0000
	Adj. R^2	.4697	.4714	S	-.1374	.1437	-.956	.3406
				I	.1409	.1348	1.045	.2979
	Effect of Adding Raters			P_1	.1844	.1298	1.420	.1578
	R^2 Change =		.0194	P_2	.0189	.1409	.134	.8933
	F Change =		1.0985	P_3	-.1979	.1425	-1.389	.1671
	F Change Sig. =		.3639					
446	R^2	.6608	.6710	TA	2.0910	.1270	16.462	.0000
	Adj. R^2	.6585	.6571	S	.0565	.1487	.380	.7044
				I	.1252	.1406	.891	.3747
	Effect of Adding Raters			P_1	.0426	.1444	.295	.7672
	R^2 Change =		.0102	P_2	.2441	.1688	1.446	.1504
	F Change =		.8793	P_3	-.0005	.1579	-.030	.9761
	F Change Sig. =		.4967					
503	R^2	.6499	.6640	TA	1.5359	.0928	16.558	.0000
	Adj. R^2	.6476	.6501	S	.1199	.1096	1.094	.2757
				I	.0842	.1196	.704	.4825
	Effect of Adding Raters			P_1	-.2339	.1121	-2.088	.0386
	R^2 Change =		.0140	P_2	-.0658	.1211	-.543	.5877
	F Change =		1.2117	P_3	.0274	.1263	.217	.8284
	F Change Sig. =		.3066					
549	R^2	.6679	.6741	TA	1.5758	.0961	16.403	.0000
	Adj. R^2	.6657	.6606	S	.0142	.1231	.115	.9085
				I	.1367	.1353	1.011	.3139
	Effect of Adding Raters			P_1	-.0834	.1291	-.646	.5193
	R^2 Change =		.0062	P_2	.1253	.1269	.987	.3251
	F Change =		.5500	P_3	-.0002	.1326	-.012	.9904
	F Change Sig. =		.7381					

Note. TA = Test Administrator; ALL = Combined effect of all raters. Task 284 not included due to lack of S, I, or P ratings. Sig T = *p* value for the T statistic.
Tests are:
154 = Perform Aircraft Support Generator Service Inspection
264 = Isolate Engine, Motor, or Generator Malfunction

Table 13 (continued)

155 = Perform Service Inspection on a Load Bank
284 = Remove and Replace Engine Fan Belts
162 = Perform Service Inspection on Hydraulic Test Stand
300 = Remove or Install Fuel Line and Fittings
179 = Perform Gas Turbine Compressor Periodic Inspection
421 = Remove or Install Hydraulic Lines
209 = Measure Resistance in AGE Electrical Systems
446 = Isolate Pneumatic System Malfunctions
215 = Perform AGE Electrical Systems Operational Checks
503 = Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance Information
238 = Splice Electrical Systems Wiring
251 = Adjust Turbine Engine Fuel System Components
549 = Inspect Vehicles for Safety of Operation
260 = Clean Motor and Generator Components

ability of the individual rater to place individuals into the five levels of job proficiency. If this value is significant then the addition of that rater accounts for a significant amount of variance in the criterion variable, HOTS.

What is interesting to note here is the drop in the adjusted R^2 for Test 154 when more predictors (i.e., raters) were added into the model. The adjusted R^2 decreases from .5495 to .5429. Prediction would seem to get worse. This indicates that the additional predictors do not buy any added variance (i.e., the added predictors are not worth the price of inclusion).

Throughout Table 13 the TA was, again, significant for every test. In only three cases did another rater achieve a significant T value: (a) Test 179 - P_2 (T=2.594, Sig T=.0104), (b) Test 215 - S (T=2.428, Sig T=.0164), and (c) Test 503 - P_1 (T=-2.088, Sig T=.0386).

Conclusions

There is, possibly, a simple reason as to why the Test Administrator (TA) would be better able to determine job proficiency. Because the TA scored the incumbent YES/NO as he or she completed the steps within a task, it might be that the TA simply rated those persons who got more steps correct at a higher level than those who got fewer correct. This may sound like too simple an answer for the difference in ratings quality. Yet because the HOTS was calculated as the-number of steps correct divided by the base score (a sum of the weights for all steps) those who correctly completed more steps would get a higher HOTS. The TA rated the incumbent directly after the performance of the task. The Supervisor (S), Incumbent (I), and Peer (P) ratings, on the other hand, were a somewhat different type of rating. The TA rating occurred in a maximal performance situation in which the incumbent was required to perform the task in a precise manner in a limited time period under the watchful (and possibly intimidating) eye of the TA. The S, I, and P raters based their ratings on recall of observed performance in their day-to-day interaction with the incumbent. While all raters had behavioral descriptors on which to base their judgments, the difference in observation methods may account for the differences in rating quality.

In summary, the ratings supplied by the S, I, and P raters provided somewhat more indirect information than that of the TA regarding the incumbent's ability to perform on the job. Thus, we believe that the TA ratings should be used to categorize the incumbents into the five levels of proficiency.

Comparison of Archival and Judgmental Standards

In comparing standards set by the archival and judgmental procedures, the first issue is which standards to use from each procedure. Recall from the discussion of the archival procedure that test administrator (TA), incumbent (I), supervisor (S), and peer (P_{1-3}) ratings were used in various combinations to establish performance standards. TA ratings were shown to more accurately classify airmen in terms of their job performance. Therefore, the standards set by using the TA ratings were selected as the prototype from the archival procedure. From the judgmental procedure, Round 2 standards were selected as the prototype. Although reliability did not change significantly between rounds, standard deviations (another form of rater agreement) decreased significantly. Also, by sharing standard setting rationales during the group discussion, SMEs presumably had more information on which to base their standards in Round 2. Thus, it is believed that Round 2 judgmental standards more accurately reflect the "truth" than Round 1 standards.

Table 14 presents standards derived from the both the archival (based on TA ratings) and the judgmental (based on Round 2 ratings) procedures for each hands-on test and across all 16 hands-on tests. Figure 8 graphically depicts the standards derived from the archival and judgmental procedures across all hands-on tests. The standards in Table 14 and Figure 8 indicate that the two standard setting procedures are producing similar results; however, potential differences were tested once again using a repeated measures ANOVA approach. The hands-on tests were treated as cases and the cutoff values as repeated measures "trials." Standard setting procedure (archival vs. judgmental) and cutoff (Marginal, Qualified, Distinguished, and Exceptional) were treated as two trials factors.

As expected, no differences were found between the standards set by the archival or judgmental procedures ($\underline{F}_{(1,15)}$=0.787, ns). A significant main effect was found for the cutoffs ($\underline{F}_{(3,45)}$=611.985, $\underline{p}$=.000), and the interaction was significant ($\underline{F}_{(3,45)}$=7.101, $\underline{p}$=.001). Planned comparisons revealed that only that the Marginal cutoff was significantly different between the two procedures ($\underline{F}_{(1,15)}$=7.002, $\underline{p}$=.018).

With the exception of the Marginal cutoff, the lack of statistically significant differences between the standards set by the archival and judgmental procedures is encouraging. Although statistically significant, the .56 difference between the archival (4.81) and judgmental (5.37) Marginal cutoffs appears to be of no practical significance. Given these results and the extensive JPM database, the archival procedure can be used to set

Table 14

Archival and Judgmental Techniques: Minimum Test Score Cutoff by Test

Test	Marginal		Qualified		Distinguished		Exceptional	
	Archival	Judgmental	Archival	Judgmental	Archival	Judgmental	Archival	Judgmental
154	6.23	6.30	7.77	7.48	8.74	8.73	9.61	9.70
155	4.48	4.98	6.32	6.50	7.93	7.97	9.03	9.12
162	6.08	5.63	7.61	7.00	8.50	8.42	9.35	9.52
179	3.51	3.85	5.58	5.35	6.56	7.00	6.62	8.20
209	5.19	5.22	7.17	6.60	8.18	8.15	9.17	9.13
215	4.75	4.77	7.79	6.47	8.85	8.05	9.69	9.17
238	6.22	5.67	7.99	7.03	9.28	8.55	9.78	9.57
251	3.57	4.35	7.21	5.92	8.45	7.83	9.33	9.17
260	4.42	5.68	7.88	7.17	8.84	8.68	9.55	9.63
264	4.81	4.65	6.23	6.00	7.84	7.57	9.40	8.78
284	5.74	6.55	7.69	7.53	8.89	8.77	9.76	9.75
300	6.19	5.85	7.61	7.18	8.73	8.57	9.57	9.57
421	2.89	5.22	4.68	6.50	6.71	8.03	8.84	9.07
446	3.28	4.80	5.61	6.03	7.17	7.93	8.90	9.18
503	4.66	5.72	6.36	7.08	8.03	8.30	9.44	9.37
549	4.94	6.62	6.79	7.82	8.26	9.02	9.38	9.83
Across all Tests	4.81	5.37	6.98	6.73	8.18	8.22	9.21	9.30

Note. Tests are:
154 = Perform Aircraft Support Generator Service Insp.
155 = Perform Service Inspection on a Load Bank
162 = Perform Service Inspection on Hydraulic Test Stand
179 = Perform Gas Turbine Compressor Periodic Inspection
209 = Measure Resistance in AGE Electrical Systems
215 = Perform AGE Electrical Systems Operational Checks
238 = Splice Electrical Systems Wiring
251 = Adjust Turbine Engine Fuel System Components
260 = Clean Motor and Generator Components
264 = Isolate Engine, Motor, or Generator Malf.
284 = Remove and Replace Engine Fan Belts
300 = Remove or Install Fuel Line and Fittings
421 = Remove or Install Hydraulic Lines
446 = Isolate Pneumatic System Malfunctions
503 = Research T.O.s for AGE Chassis, Enclosure, and Drive Maintenance Information
549 = Inspect Vehicles for Safety of Operation

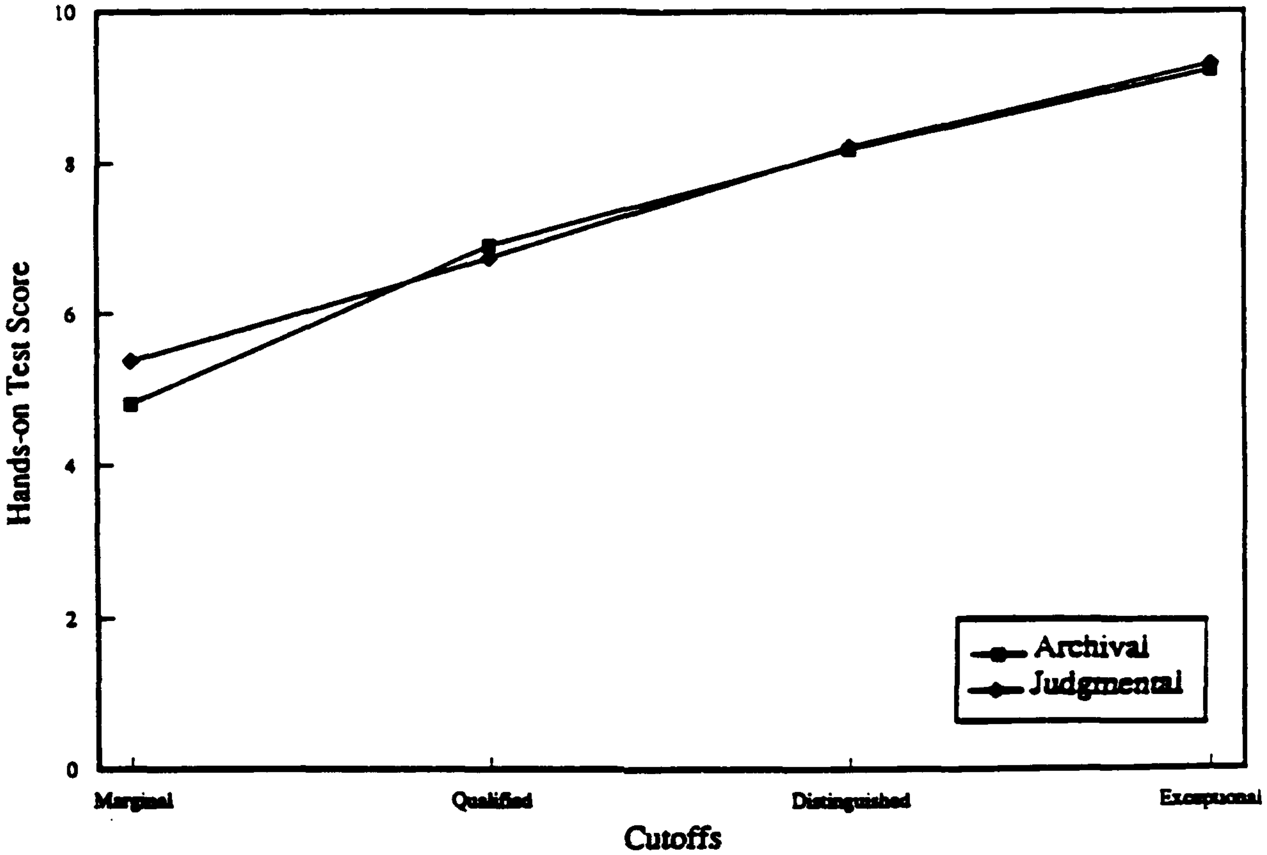

Figure 8. Archival and judgmental techniques: Standards across all hands-on tests.

standards on hands-on tests for the remaining seven AFSs under study in the JPM Project as well as for the interview tests. Using the archival procedure would result in considerable financial savings to the Air Force. One drawback of the archival procedure, however, is that laypeople tend to have less confidence in the standards set by such a procedure than in standards set by a panel of experts (Poggio, 1984). Laypeople tend to be confused by and have little faith in the statistical procedures used to yield standards via an archival approach. Because the two procedures resulted in analogous standards, the results of the present study can be used to help convince users of the authenticity of standards established with an archival procedure.

The archival procedure as described here cannot be used to establish minimum standards for the writter, job knowledge tests. In the present study, the archival procedure used technical proficiency (TP) ratings provided by the TA for each of the 16 hands-on tests to set standards on those tests. That is, TP-TA ratings for Test 154--Perform an Aircraft Support Generator Service Inspection were used to set standards on Test 154. Because the TP-TA ratings are task/test specific, they are not appropriate for setting standards on the written tests. Taken together, the written tests cover more aspects of the technical portion of the airman's job than do the hands-on tests. TP ratings are not available from any source (i.e., TA, S, I, or P) for each of the tasks assessed by the written tests. Therefore, TP ratings cannot be used to set standards on the written tests. The GTP ratings, which would be appropriate for setting standards on the written tests, cannot be used because they are unreliable (Hedge & Ringenbach, 1989). Thus, a judgmental procedure is required for establishing minimum performance standards on the written tests.

Topics for Further Consideration

The present study successfully established minimum job performance standards for the AGE hands-on tests. The archival and judgmental procedures described here can be used to set minimum standards for hands-on tests in other AFSs as well as to set standards for the interview tests. The standards established in the present study can be used for selection and classification purposes as well as to assess training.

Selection and Classification

In terms of selection and classification, the standards can be used in several ways, including but not limited to (a) the identification of personnel who may need additional training, (b) a part of a skills improvement motivation system, and (c) the establishment of selection cutoffs. However, the study raises at least two selection and classification issues which are as yet unresolved. One issue concerns the identification of a dichotomous cutoff. A second issue concerns the method of combining standards. These issues are discussed briefly below.

Recall from the introduction that establishing minimum performance standards is merely a step in the overall goal of linking job performance standards to enlistment standards. The linkage process is complex and beyond the scope of the present study. However, one aspect of the process closely related to the present study concerns the identification of a dichotomous cutoff. Although the Air Force, or any other organization, would like to select only those individuals who would be expected to be either Exceptional or Distinguished performers, such a restrictive selection cutoff is unrealistic. Yet, intentionally selecting individuals who would likely be classified as Unqualified performers is also impractical. A more realistic cutoff would result in the selection of individuals who would be expected to be either at least Marginal or at least Qualified performers. The question then becomes what is the lowest proficiency level the Air Force is willing to accept: Marginal or Qualified.

Practical and economic considerations will drive the decision to designate either Marginal or Qualified as the definitive selection cutoff. Of practical consideration is the fact that (a) the population of 18-year olds, which constitutes the Air Force's applicable labor pool, is shrinking and (b) the aptitude levels of those individuals are also decreasing (Ramsberger & Means, 1987). Thus, the recruiting and training costs of maintaining an Air Force with sufficient defense capabilities are likely to increase in the near future. In addition to these market-driven costs, the Air Force will incur additional costs depending on the selection cutoff established. By definition, Marginal personnel will cost the Air Force more in terms of training dollars. Marginal personnel may also be likely to leave the Air Force before completing their enlistment term thus preventing the Air Force from fully realizing its additional training investment. Qualified personnel, compared to Marginal personnel may require additional recruitment expenditures. In identifying a definitive selection cutoff, decision makers must weigh the additional training costs associated with selecting Marginal personnel against the additional recruitment dollars necessary to access and retain Qualified personnel.

The second unresolved issue concerns the method of combining standards. The present study used a compensatory model to combine standards for the 16 hands-on tests to yield an overall hands-on performance standard. By averaging the standards set on the 16 hands-on tests, we allowed good performance on one test to compensate for poor performance on another test. Given the information available at this time, a compensatory model was the only way to approach the issue of combining test standards. However, a compensatory model may not be appropriate. It may be that a multiple hurdle approach best describes the proper procedure for aggregating standards. Perhaps high performance on one test cannot compensate for poor performance on other tests. Further research is needed to determine the proper method for aggregating standards. In addition to the appropriate method for aggregating standards within a job performance measure (i.e., hands-on tests), research is needed to determine the relevant method for aggregating standards across assessments (i.e., hands-on tests, written tests, and interview tests).

Training Evaluation

In addition to addressing selection and classification concerns, the standards established in the present study could be used to augment the Air Force's Field Evaluation System. The Field Evaluation System is used to assess and, if necessary, to revise the technical training curriculum. The system currently uses questionnaires and field visits to evaluate technical school training. Although on-the-job training (OJT) is a critical component of Air Force training, the Field Evaluation System does not include an evaluation of OJT. Administering the hands-on tests to recent graduates and more experienced airmen during field visits, classifying incumbents into the five proficiency levels defined in the present study, and comparing the performance of recent graduates to that of more experienced airmen would allow for comparisons of technical training and OJT. Decisions could be made concerning whether tasks should be taught in technical school or in OJT, whether to allocate more or less training time to a given task in technical school versus in OJT, etc.

Ford & Sego (1990) developed a conceptual model linking training evaluation to training needs assessment. Their model outlined five major purposes of conducting training evaluation, the questions asked for each purpose, and the types of information needed to answer the questions. The purposes for conducting training evaluation were to: (1) determine the content validity of the training program, (2) examine training efficiency, (3) determine training validity, (4) determine transfer validity, and (5) determine the predictive validity of a training program.

Additional research is needed to continue the development of this training evaluation typology for the Air Force. Specifically, the information that is available to conduct content validity, training validity, transfer validity, and predictive validity evaluations should be reviewed and documented. This effort should begin with an extensive exploration and, if necessary, development of operational sources of information. The domains of interest ar(those domains obtained when the sources and types of information in the model are combined (i.e., job content domain, training content domain, job performance domain, and training performance domain).

In summary, the model provides a conceptual framework for describing the information available to answer training evaluation questions. Ford & Sego (1990) documented some of the types of evaluative information gathered for each of the four domains mentioned above. This produced an analysis of the issues relevant to training evaluation and a critique of the existing training evaluation system on the issues of content validity and training efficiency. This analysis and critique should be expanded to the issues associated with training validity, transfer validity, and predictive validity for both the job domain and the training domain. This additional analysis and critique will provide the information necessary to begin the development of an enhanced training evaluation system (both resident and on the job) for the Air Force.

References

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 36, 45-50.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. Journal of Applied Psychology, 60, 556-560.

Brennan, R. L., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. Applied Psychological Measurement, 4, 219-240.

Buck, L. S. (1977). Guide to the setting of appropriate cutting scores for written tests: A summary of the concerns and procedures (Technical Memorandum 77-4). Washington, DC: Personnel Research and Development Center, United States Civil Service Commission.

Cantor, J. A. (1989). A validation of Ebel's method for performance standard setting through its application with comparison approaches to a selected criterion-referenced test. Educational and Psychological Measurement, 49, 709-721.

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

Cross, L. H., Impara, J. C., Frary, R. B., & Jaeger, R. M. (1984). A comparison of three methods for establishing minimum standards on the National Teacher Examinations. Journal of Educational Measurement, 21, 113-130.

Elder, B. L., & Hansen, L. A. (1991). Examining operational measures of performance and developing methods for determining competency levels for the Air Force Job Performance Measurement System: Performance measures matrix (Final Report FR-PRD-91-05). Alexandria, VA: Human Resources Research Organization.

Ford, J. K. & Sego, D. (1990). Linking training evaluation to training needs assessment: A conceptual model (AFHRL-TP-90-69). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

Fotouhi, C. H., Mosher, G. P., & McCloy, R. A. (1990). Examining operational measures of performance and developing methods for determining competency levels for the Air Force Job Performance Measurement System: Literature review and methods (Interim Report IR-PRD-90-08). Alexandria, VA: Human Resources Research Organization.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. American Psychologist, 18, 519-521.

Halpin, G., & Halpin, G. (1987). An analysis of the reliability and validity of procedures for setting minimum competency standards. Educational and Psychological Measurement, 47, 977-983.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. Educational and Psychological Measurement, 43, 185-196.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Burk (Ed.) Criterion referenced measurement: The state of the art. Baltimore, MD: Johns Hopkins Press.

Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 15, 277-290.

Hedge, J. W., & Ringenbach, K. L. (1989). An assessment of rating form quality across eight Air Force Specialties. Report prepared for the Air Force Human Resources Laboratory, Training Systems Division, Brooks AFB, TX.

Jaeger, R. M. (1976). Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 18, 22-27.

Jaeger, R. M., & Busch, J. C. (1984). The effects of a delphi modification of the Angoff-Jaeger standard-setting procedure on standards recommended for the National Teacher Examinations. Paper presented at the joint annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA. (ERIC Document 246 091).

Jaeger, R. M., & Keller-McNulty, S. (1986, July). Procedures for eliciting and using judgments of the value of observed behaviors on military job performance tests. Prepared for the Committee on the Performance of Military Personnel, Commission on Behavioral and Social Sciences and Education, National Research Council/National Academy of Sciences.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. Journal of Educational Measurement, 17, 167-178.

Latham, G. P., Wexley, K. N., & Pursell, E. D. (1975). Training managers to minimize rating errors in the observation of behavior. Journal of Applied Psychology, 60, 550-555.

Lipscomb, M. S., & Dickinson, T. L. (1987). Test content selection. In H. G. Baker & G. J. Laabs (Eds.). Proceedings of Department of Defense/Educational Testing Service Conference on Job Performance Measurement Technologies. San Diego, CA: Navy Personnel Research and Development Center.

Livingston, S. A., & Kastrinos, W. (1982). A study of the reliability of Nedelsky's method for choosing a passing score (Report No. ETS-RR-82-6). Princeton, NJ: Educational Testing Service. (ERIC Document No. ED 218 361).

Mills, C. N. (1983). A comparison of three methods of establishing cutoff scores on criterion-referenced tests. Journal of Educational Measurement, 20, 283-292.

Norcini, J. J., Lipner, R. S., Langdon, L. O., & Strecker, C. A. (1987). A comparison of three variations on a standard-setting method. Journal of Educational Measurement, 24, 56-64.

Office of the Assistant Secretary of Defense (Force Management and Personnel). (1989). Joint service efforts to link enlistment standards to job performance: Recruit quality and military readiness. Report to the House Committee on Appropriations, Washington, DC.

Peterson, N. G., Owens-Kurtz, C., Hoffman, R. G., Arabian, J. M., & Whetzel, D. L. (1989). Army synthetic validation project: Report of Phase II results (Volume I) (ARI Technical Report). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Plake, B. S., & Melican, G. J. (1989). Effects of item content in initial judge consistency of expert judgments via the Nedelsky standard setting method. Educational and Psychological Measurement, 49, 45-51.

Poggio, J. P. (1984). Practical considerations when setting test standards: A look at the process used in Kansas. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Poggio, J. P., Glassnap, D. R., & Eros, D. S. (1981, April). An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Pulakos, E. D., & Borman, W. C. (1985). Development and field test of the Army-wide rating scales and rater orientation and training program (ARI Technical Report 716). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.

Pulakos, E., Wise, L., Arabian, J., Heon, S., & Delaplane, S. K. (1989). A review of procedures for setting job performance standards. Washington, DC: U.S. Army Research Institute for the Behavioral and Social Sciences.

Ramsberger, P. F., & Means, B. (1987). Military performance of low aptitude recruits: A reexamination of data from Project 100,000 and the ASVAB misnorming period (FR-PRD-87-31). Alexandria, VA: Human Resources Research Organization.

Shavelson, R. J., Webb, N. M., & Rowley, G. L. (1989). Generalizability theory. American Psychologist, 44, 922-932.

Shephard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4, 447-467.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. Journal of Educational Measurement, 17, 229-235.

Wigdor, A. K., & Green, B. F. (eds.). (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, DC: National Academy Press.

Wise, L. L., Peterson, N. G., Hoffman, R. G., Campbell, J. P., & Arabian, J. M. (1990). Army Synthetic Validity Project: Report of Phase III results: Volume I. (Draft Report). Washington, DC: American Institutes for Research.

Wood, R., & Power, C. (1987). Aspects of the competence-performance distinction: Educational, psychological, and measurement issues. Journal of Curriculum Studies, 19, 409-424.