(12)

AFHRL-TR-86-22

AIR FORCE HUMAN RESOURCES LABORATORY

ON-THE-JOB TRAINING: DEVELOPMENT AND ASSESSMENT OF A METHODOLOGY FOR GENERATING TASK PROFICIENCY EVALUATION INSTRUMENTS

Ronnie Warm
J. Thomas Roth
Jean A. Fitzpatrick

Applied Science Associates, Inc.
P.O. Box 1072
Butler, Pennsylvania 16003

TRAINING SYSTEMS DIVISION
Lowry Air Force Base, Colorado 80230-5000

September 1986

Final Report for Period January - October 1985

Approved for public release; distribution is unlimited.

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

NOTICE

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

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

This report has been reviewed and is approved for publication.

GERALD S. WALKER
Contract Monitor

JOSEPH Y. YASUTAKE, Technical Director
Training Systems Division

DENNIS W. JARVI, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release; distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFHRL-TR-86-22

6a. NAME OF PERFORMING ORGANIZATION
Applied Science Associates, Inc

6b. OFFICE SYMBOL *(If applicable)*

7a. NAME OF MONITORING ORGANIZATION
Training Systems Division

6c. ADDRESS *(City, State, and ZIP Code)*
P.O. Box 1072
Butler, Pennsylvania 16003

7b. ADDRESS *(City, State, and ZIP Code)*
Air Force Human Resources Laboratory
Lowry Air Force Base, Colorado 80230-5000

8a. NAME OF FUNDING / SPONSORING ORGANIZATION
Air Force Human Resources Laboratory

8b. OFFICE SYMBOL *(If applicable)*
HQ AFHRL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
F33615-82-C-0004

8c. ADDRESS *(City, State, and ZIP Code)*
Brooks Air Force Base, Texas 78235-5601

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
62205F	1121	09	10

11. TITLE *(Include Security Classification)*
On-the-Job Training: Development and Assessment of a Methodology for Generating Task Proficiency Evaluation Instruments

12. PERSONAL AUTHOR(S)
Warm, R.; Roth, J.T; Fitzpartick, J.A.

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM Jan 85 TO Oct 85

14. DATE OF REPORT *(Year, Month, Day)*
September 1986

15. PAGE COUNT
138

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

assessment instrument OJT evaluation
assessment methodology performance assessment
on-the-job training performance evaluation (Continued)

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

This document describes the development and assessment of a methodology for generating on-the-job training (OJT) task proficiency assessment instruments. The Task Evaluation Form (TEF) development procedures were derived to address previously identified deficiencies in the evaluation of OJT task proficiency. The TEF development procedures allow subject-matter experts (SMEs), without experience in assessment methodology, to construct task proficiency assessment instruments which can be used by OJT supervisors to assess trainee proficiency at specific tasks. The reliability and validity of the procedures and of the forms resulting from procedural application were examined in two Air Force career fields (Aircraft Maintenance and Security Police/Law Enforcement). The results of the procedural assessment indicated that SMEs could reliably apply the development procedures to construct TEFs which accurately and completely depict the critical aspects of task performance. The operational assessment demonstrated that evaluator subjectivity in the assignment of end scores and pass/fail decisions is minimized when TEFs are used. In addition, it was shown that evaluations conducted with TEFs provide end scores and pass/fail decisions which accurately reflect the number and types of errors aetected in task performance. The combined results of the two assessments demonstrate the

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION

22a. NAME OF RESPONSIBLE INDIVIDUAL
Nancy A. Perrigo, Chief, STINFO Office

22b. TELEPHONE *(Include Area Code)*
(512) 536-3877

22c. OFFICE SYMBOL
AFHRL/TSR

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

18. (Concluded)

performance standards
task evaluation
task performance
task proficiency

19. (Concluded)

potential of the TEFs in the OJT environment. This document includes an overview of the theoretical approach, and detailed descriptions of the assessment methodologies and results. The final section of this document contains recommendations for the integration of the TEF development procedures and the TEFs into the OJT environment.

SUMMARY

The ability to accurately and objectively assess an individual's level of performance on the job is important to Air Force systems for personnel selection, assignment, training, and utilization. Currently, the on-the-job training (OJT) supervisor is responsible for the evaluation of an individual trainee's task proficiency. Concern has been expressed within the Air Force training community regarding the variability which exists among supervisors in the evaluation of OJT task proficiency. In particular, there exists a lack of standardization with regard to the construction of assessment instruments, administration of performance evaluations, the scoring of results, and providing feedback.

The Task Evaluation Form OJT task proficiency assessment system (TEF System) which was developed and evaluated by Applied Science Associates, Inc. (ASA), enables subject-matter experts (SMEs), without experience or training in assessment methodology, to develop instruments which can be used to assess OJT task proficiency. SMEs apply prescribed development procedures to a specific task in order to generate a Task Evaluation Form (TEF). Evaluators use the TEFs to conduct evaluations of OJT task performance. The TEF development procedures combine logical analysis and a critical incident technique, resulting in the identification and description of task elements which are critical to successful task performance. Potential performer actions and outcomes are divided into evaluation areas: Time/Speed; End Product; Sequence-Following; Safety; and Tools, Equipment, and Materials Use. The TEF development procedures guide the developer in determining which evaluation areas are critical to successful task performance. In addition to guiding the developer in identifying and describing critical aspects of task performance, the TEF development procedures also provide instructions for creating an evaluation scenario and developing a chart for scoring task performance.

The reliability and validity of the TEF development procedures (procedural assessment), and of the TEFs resulting from their application (operational assessment), were examined. The results of the procedural and operational assessments indicate that the TEF system can be utilized within the OJT environment. SMEs are able to apply the TEF development procedures to generate assessment instruments. The resulting TEFs contain: a description of the evaluation scenario, critical areas of performance for evaluation, designation of the steps or events where each area should be evaluated, standards of performance for each identified event, and criteria for scoring observed performance and assigning a pass/fail decision. When supervisors use TEFs to assess task proficiency, results are derived which consistently reflect observed performance. Most importantly, the TEF system provides evaluation results which are useful and meaningful within the OJT environment. The standardization of the evaluation situation allows for tracking an individual's performance, as well as making comparisons between performers possible. In addition, the evaluation results actually summarize task performance by providing a separate score for each applicable evaluation area. These scores can be used to provide objective feedback to the performer and/or trainer. The TEF system has demonstrated potential for several uses in the OJT environment; e.g., determining an individual's performance level and acceptance for task certification, tracking individual or unit level performance, comparing performers, identifying specific deficiencies in an individual's OJT, and identifying OJT deficiencies on a unit level.

Captain Richard Dineen	AFHRL/ID, Lowry AFB
MSgt Robert Aust	4235 Training Squadron, Carswell AFB
Captain Neil Breckenridge	4235 Training Squadron, Carswell AFB
TSgt Gilbert Elmy	4235 Training Squadron, Carswell AFB
MSgt Goolsby	4235 Training Squadron, Carswell AFB
MSgt Thadous J. Magwood	HQ AFOSP, Kirtland AFB
MSgt Larry R. Schmidt	HQ AFOSP, Kirtland AFB
CMSgt David H. Smith	1606 SPG, Kirtland AFB
Captain Sherry D. Webb	HQ AFOSP, Kirtland AFB
Major Martin Costellic	AFHRL/OL-AK, Bergstrom AFB
Major Larry Johnston	AFHRL/OL-AK, Bergstrom AFB
Major Gary Martilla	CBPO, Bergstrom AFB
Colonel Charles Daye	CBPO, Bergstrom AFB

PREFACE

This document was prepared by Applied Science Associates, Inc. (ASA), Valencia, Pennsylvania, under Air Force Contract Number F33615-82-C-0004. Ms. Ronnie E. Warm was the Project Scientist and Director. The project was sponsored by the Air Force Human Resources Laboratory, Training Systems Division, Lowry Air Force Base, Colorado. The Contract Monitor was Mr. Gerald S. Walker. Captain Richard Dineen served as the Technical Contract Monitor until his retirement in July 1984, at which time Major Martin Costellic became the Technical Contract Monitor.

This study is one of a series of related studies under the Program: Systems Integration, Transition, and Technical Support. The objective of this program is to provide support for the Air Force's Advanced On-the-Job Training System (AOTS). Task proficiency assessment instruments are necessary for assessing training on an individual and unit level within the AOTS. However, techniques for performing reliable, valid, and standardized task proficiency evaluations do not currently exist. The objective of this effort was to devise a methodology for the development of on-the-job training (OJT) task proficiency assessment instruments. Procedures (Task Evaluation Form development procedures) were derived which allow subject-matter experts (SMEs) to construct task proficiency assessment instruments. The instruments resulting from application of the procedures (Task Evaluation Forms) can be used by OJT supervisors to assess trainee proficiency at specific tasks. This document describes the activities related to the derivation and assessment of the Task Evaluation Form (TEF) development procedures and the instruments resulting from their application. The actual development procedures are contained in two separate documents: Task Evaluation Form: Development Procedures for Maintenance and Equipment-Oriented Tasks; and Task Evaluation Form: Development Procedures for Non-Equipment-Oriented Tasks (AFHRL Technical Papers 85-55 and 85-56, respectively).

The authors wish to acknowledge the assistance and cooperation of the many individuals who contributed to the completion of this document. From ASA, appreciation is expressed to: Rohn J. Hritz and George R. Purifoy, Jr. for their contributions to the initial derivation of the Task Evaluation Form (TEF) development procedures, and their support and ideas throughout the project; Ms. Lisa I. Thocher for her assistance in the data collection and analysis; and the Project Secretary, Mrs. Tammy Mowry. Special thanks are due to Mr. Gerald S. Walker, the Air Force Contract Monitor.

This project would not have been possible without the cooperation of all of the Air Force SMEs and project coordinators. The authors are especially grateful to the following:

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

TABLE OF CONTENTS (Continued)

LIST OF TABLES

words, the assessment instrument focuses the supervisor's attention upon specific aspects of task performance. The instrument details for the supervisor what should be evaluated. Standardization of evaluation content is necessary in order to compare evaluation results obtained by different evaluators. Currently, there are no standard formats for instrument development. Often, assessment instruments or evaluation forms are not provided. The content of the evaluation is left entirely to the individual supervisor's discretion. Evaluator subjectivity is introduced into the evaluation process when different evaluators focus on different aspects of task performance.

When assessment instruments are provided, they are usually in the form of checklists. The checklists contain lists of items which must be completed by the performer. However, they do not include standards of performance; i.e., descriptions of acceptable behaviors or outcomes. The decision as to what qualifies as acceptable task performance is left to the individual supervisor. Thus, even though two supervisors may focus on the same aspects of performance (i.e., those aspects included on the checklist), they will not always agree about what qualifies as acceptable completion of those items.

In summary, the lack of a standard instrument for evaluating task proficiency leads to evaluator subjectivity regarding what aspects of task performance are evaluated. Even when evaluator agreement exists regarding what is to be evaluated, differences may occur regarding what standards must be met. Thus, there is no guarantee that two different performers or the same performer evaluated by two different supervisors would be required to demonstrate the same level of task proficiency.

1.1.2 Administration of Task Proficiency Assessment

In addition to focusing on different aspects of the task, evaluators tend to conduct job performance evaluations in different ways. In order to compare evaluation results obtained by different evaluators, standardization is necessary. However, evaluator differences occur frequently. Some of the more common evaluator differences are mentioned below.

1. Some supervisors use the evaluation situation as a training experience for the performer. These supervisors have a tendency to immediately correct the performer when errors in performance occur. This immediate feedback confounds the evaluation process and results.

2. The evaluation situation is often used as an opportunity to test "systems knowledge" through questions to the performer. During task performance, some evaluators assess both knowledge and performance ability, whereas other supervisors assess only performance ability.

1.0 INTRODUCTION

The ability to objectively and accurately assess an individual's level of proficiency is critical to that individual's training and utilization. During Air Force on-the-job training (OJT), the OJT supervisor is required to assess each airman's proficiency. The assessment of proficiency is typically divided into two types of evaluation: the evaluation of job knowledge and the evaluation of job performance. Job knowledge is usually assessed through written tests. These tests are standardized such that every airman (in the same career field) is tested on the same subjects, in the same way. In addition, a standardized scoring routine is used to score written tests. Thus, for the assessment of job knowledge, the test content, administration, and scoring are standardized. Standardization is important in order for the evaluation results to be used in meaningful ways; i.e., to compare performers tested by different evaluators or to track an individual's performance over time.

Job performance is typically assessed by evaluating the performance of individual tasks within a specialty area. In other words, the individual being evaluated must demonstrate proficiency in performing selected tasks from his/her career field. However, the assessment of task performance/proficiency is not as standardized as the assessment of job knowledge.

Standardization is lacking in the areas of instrument development, administration, and scoring. Thus, variability exists regarding what is evaluated, how evaluations are conducted, and how observed task performance is scored. This lack of standardization in the evaluation of task proficiency makes it difficult to use the evaluation results to compare performers or to track an individual's progress over time.

This project was initiated to address the aforementioned issues in the evaluation of OJT task proficiency. In this section, the existing problems in the evaluation of task proficiency are discussed, a summary of the approach is presented, and the organization of the rest of the report is described.

1.1 Existing Problems in the Evaluation of OJT Task Proficiency

The central area of deficiency in the evaluation of OJT task proficiency is the lack of standardization. This lack of standardization as it relates to several areas of assessment methodology is discussed below.

1.1.1 Assessment Instrument Development

The particular assessment instrument used to evaluate task proficiency determines the actual content of the evaluation. In other

3. Supervisors allow varying amounts of assistance during task performance.

4. For many tasks, the degree of difficulty and/or performance requirements vary as environmental conditions are altered. However, evaluators tend to select different environmental conditions under which to conduct an evaluation.

When one or more of these evaluator differences exist, it becomes impossible to compare the evaluation results obtained by two different evaluators.

1.1.3 Scoring the Evaluation

The third area of supervisor variability is in the derivation of an end score and assignment of a pass/fail decision. Currently, there is no guidance regarding how observed errors in performance should affect the end score or pass/fail decision. There are no specific criteria to define the number and types of errors which are permitted in task performance. Conversely, there are no criteria to define the number and types of errors which should result in failure of task performance. Thus, an evaluator has no means of arriving at a meaningful end score. In addition, there is no guarantee that two supervisors who observe the same task performance will derive identical end scores and pass/fail decisions.

1.1.4 Feedback

Feedback regarding deficiencies in task performance is important on an individual and unit level. For example, repeated individual deficiencies in the same area may indicate a need for remedial training. On the other hand, performance deficiencies in a particular area on a unit level may indicate deficiencies in the OJT program.

Currently, there is no standard means for providing specific feedback regarding task performance. Specificity of feedback is important to an evaluation system if steps are to be taken to improve performance.

1.2 Approach Overview

The existing problems in the evaluation of OJT task proficiency were addressed by Applied Science Associates, Inc. (ASA) in a task ordering contract with the Air Force Human Resources Laboratory. The overall approach involved several phases. During the first phase, an examination of the OJT environment was conducted and a list of requirements for an effective OJT assessment system was derived. Next, a task proficiency assessment system, titled the Task Evaluation Form

(TEF) system was developed. The usefulness and applicability of the TEF system in the OJT environment was assessed in the last phase of the project.

This report contains details of each phase of the project. In addition, conclusions and recommendations are provided.

1.3 Overview of This Report

The remainder of this report is divided into six sections as follows:

1. Section 2.0 includes a discussion of relevant issues in the development of an OJT task proficiency assessment system and a list of requirements for an effective OJT task proficiency assessment system. In addition, an overview of the Task Evaluation Form development procedures is presented in Section 2.0.

2. The methodology and results of the procedural reliability and validity assessment are described in Section 3.0.

3. In Section 4.0, the details of operational reliability and validity assessment are provided.

4. Section 5.0 includes a description of the application of the TEF development procedures in the Personnel career field.

5. A discussion of the TEF methodology in relation to the requirements set forth in Section 2.0 is presented in Section 6.0.

6. In Section 7.0, recommendations for the use of the TEF system are discussed.

2.0 THEORETICAL APPROACH

There are many basic issues and considerations which affect the development and use of task proficiency assessment instruments in a specific environment. Before procedures for developing OJT task proficiency assessment instruments could be recommended, it was necessary to obtain information about how the development procedures and the resulting instruments would be used in the OJT environment.

Information about current evaluation procedures in the OJT environment was gathered from various sources. Assumptions about the development and use of OJT assessment instruments were made based upon the information obtained.

In this section, assumptions about the use of assessment instruments in the OJT environment and a list of requirements for an effective OJT task proficiency assessment system are presented.

2.1 Use of Assessment Instruments in the OJT Environment

In order to develop effective assessment instrument development procedures, it is important to know how the assessment instruments should fit in the intended environment (OJT).

An assessment of the OJT environment was conducted in order to answer the following questions:

How will the assessment instruments be used?

Who will develop the assessment instruments?

What method(s) of assessment will be used?

For what task types will assessment• instruments be developed?

Information was gathered from the following sources:

Review of technical documents on OJT operational and administrative procedures.

Interviews with OJT personnel.

Review of existing task databases.

Review and analysis of existing assessment instrument develcpment procedures.

A set of specific assumptions was derived from the answers to the aforementioned questions. Each question and the resulting assumptions are discussed below.

2.1.1 How Will the Assessment Instruments be Used?

The purpose of an assessment instrument influences the types of evaluation information collected. Thus, it was important to know the purpose for which assessment results were intended before an assessment instrument could be designed.

It was conceivable that OJT task proficiency instruments would be used for a multitude of purposes. For instance:

1. To determine the proficiency level of individual airmen for certification and/or promotion.

2. To determine individual deficiencies in task proficiency.

3. To determine the state of readiness of a given unit.

4. To determine the training effectiveness of an OJT program (i.e., to determine OJT program deficiencies rather than individual deficiencies).

It is clear that the specific purpose influences the amount of detail that the assessment instrument must provide. For instance, more detailed information is necessary to determine deficiencies on an individual or OJT program level than is required to make decisions about certification or unit readiness.

Since the specific purpose of the instruments was not clear at the onset of the project, it was necessary to design an assessment instrument which would generate the highest possible level of detail. This would allow the instruments to be used for several purposes.

2.1.2 Who Will Develop the Assessment Instrument?

The characteristics of the target population will influence the nature and complexity of the development procedures.

Several potential target groups were identified. It was possible that the following groups might be tasked with assessment instrument development.

1. OJT supervisors.

2. Flightline or job supervisors.

3. Training development personnel.

4. A special group - specifically formed for the purpose of developing assessment instruments.

Thus, it was necessary to design assessment instrument development procedures which could be applied by individuals from any of the above groups. An iterative approach to the design of the development procedures was chosen. Procedures were developed and revised following trial applications by members from each potential target group.

2.1.3 What Method of Assessment Will be Used?

Assessment instruments should reflect the method of assessment which should be used. Different types of information are included on assessment instruments, depending upon the intended evaluation method.

Several different methods are currently used to assess OJT task proficiency. These methods are:

1. Direct observation of the performer performing tasks in the job environment.

2. Direct observation of the performer performing the task in a controlled environment; i.e., a rigged task scenario.

3. Direct observation of the performer performing on a simulator or a trainer.

4. Paper-and-pencil testing of job knowledge. This approach is more appropriate for testing job knowledge than for testing job performance.

The determination of the most appropriate evaluation method depends upon the task for which the instrument is developed. The development procedures must allow the development of assessment instruments which could accommodate any of the direct observation methods. In addition, the development procedures must include guidance for the developer in selecting the appropriate evaluation method.

2.1.4 For What Task Types Will the Assessment Instruments be Developed?

An examination of the OJT environment revealed the broad range of task types which are trained in Air Force OJT.

Due to this diverse nature of tasks, task-specific development procedures were deemed impractical and certainly not feasible. Thus, it was determined that the development procedures must be generic enough to apply to a broad range of task types. However, it was also necessary that the development procedures allow the depiction of the important dimensions of any task performance.

In order to obtain the necessary balance between generalizability and specificity of procedural application, a population of OJT task performance dimensions must be identified. The decisions regarding the inclusion of individual performance dimensions must be task specific. The development procedures must include criteria for making those decisions.

2.2 Summary of Assumptions

The following assumptions were used to guide the initial design of the assessment instrument development procedures:

1. The results of evaluations with the assessment instruments may be used to make decisions about individual certification, individual deficiencies, the effectiveness of an OJT program, or the state of readiness of a given unit.

2. The assessment instruments will be developed by technically qualified military personnel who are inexperienced in performance assessment methodology.

3. The assessment instrument will be designed to conduct "hands-on" or "over-the-shoulder" evaluations in the OJT environment, and the development procedures will include guidelines for selecting a specific "over-the-shoulder" method.

4. The assessment instrument development procedures will be capable of application to a wide variety of task types.

It should be noted here that the following issues were not addressed in this project:

1. The selection of tasks for which assessment instruments should be developed.

2. The selection of tasks for which evaluations should be conducted.

2.3 General Requirements for an Effective OJT Task Proficiency Assessment System

Based upon the existing problems in the evaluation of OJT task proficiency and the assumptions about the use of OJT assessment instruments, a list of requirements for an effective OJT assessment system was derived.

Two types of requirements were derived: general requirements related to the application of the assessment instrument development procedures (referred to as procedural requirements), and requirements related to the use of the assessment instruments (operational requirements).

2.3.1 Procedural Requirements

An effective OJT task proficiency assessment system must provide procedures for instrument construction which have the following characteristics:

1. Validity. The development procedures must result in the identification of those aspects of task performance which are directly related to successful task performance.

2. Reliability. The development procedures must be consistently applied by two or more users. In other words, two or more users applying the procedures to the same task should identify the same aspects of task performance.

3. Utility. The development procedures must have the capability of application by Air Force users. Although a specific target population has not been identified, it is expected that the users will not have training or experience in assessment methodology.

4. Generalizability. The assessment instrument development procedures must be applicable to tasks from all specialty areas including both maintenance and non-maintenance specialty areas.

2.3.2 Operational Requirements

The use of the assessment instruments must provide evaluation results which have the following characteristics:

Validity. The results must accurately reflect the number and types of errors which occur in task performance.

Reliability. Two or more evaluators, observing the same task performance, should derive identical results.

Standardization. Two or more evaluators, conducting an evaluation of the same task, should conduct the evaluation in the same manner.

Utility. The results of evaluations must provide direct feedback regarding specific deficiencies in task performance. In addition, varying levels of detail of feedback must be provided so that the user can select the appropriate level for the intended purpose.

2.4 Performance Evaluation Guides (PEGs)

The procedure or methodology described in this report was partially derived from a previous Air Force effort. Before presenting an overview of the assessment instrument development procedures, it will be beneficial to briefly review the previous methodology.

The previous Air Force effort on performance assessment, referred to as Performance Evaluation Guides (PEGs), involved the design of procedures for developing task performance assessment instruments. The method used was described as a "loosely structured critical incident technique." The PEGs development procedures began by having developers decompose tasks into their component steps. The task steps were then reviewed to identify potentially "critical errors." Only those steps that contain "critical" errors are to be assessed. Errors are identified as either process (key steps) errors or product (end results) errors.

Although the PEGs procedures are relatively easy to follow, the effectiveness of the process is questionable. The important issues are whether or not the process actually identifies critical measures of successful task performance (validity), and whether or not the procedures can be consistently applied (reliability). For the following reasons, it is highly unlikely that the PEGs process is both valid and reliable:

1. The procedures provide little guidance for identifying potentially critical errors and determining measures of successful task performance.

2. The critical incident technique is an acceptable method for identifying task components that are important to measure. However, PEGs use the technique in a limited way. Critical errors are identified for each step, but critical task performance is never described. Thus, a PEGs instrument does not reflect what behaviors are to be evaluated.

3. A PEGs instrument is basically a checklist, limited to key steps. PEGs do not identify what dimensions of task performance are important to evaluate, such as safety, time, use of tools, or features of task output; nor do they identify performance standards.

4. PEGs provide only GO/NO-GO decisions. A GO/NO-GO dichotomy may not be appropriate for all task evaluations. Descriptive information about task performance is necessary to provide effective feedback about performance.

2.5 Overview of Task Evaluation Form Development Procedures

The method for developing performance assessment instruments [the Task Evaluation Form (TEF) development procedures] described in this document expands the PEGs methodology to combine a logical analysis approach with the critical incident technique. The critical incident technique is used to identify steps and events that are important to evaluate. The logical analysis approach is used to guide the developer in making decisions about how the evaluation should be conducted and what behaviors or outcomes reflect successful task performance. Thus, the critical incident technique used in this method initially focuses the developer's attention on the consequences of not successfully completing events in the task. Once critical events are identified, the logical analysis approach is used to guide the developer in making decisions about what performer actions and outcomes actually reflect successful task performance.

Possible performer actions and outcomes are divided into five evaluation areas (the evaluation of: Time/Speed, Sequence-Following, End Product, Safety, and Tools, Equipment, and Materials Use). The developer is guided in determining which areas should be evaluated during task performance. The developer considers every step or event in the task and identifies the steps or events in the task where a particular evaluation area should be evaluated. Once critical task events or steps have been identified, instructions are provided for describing successful task performance. These descriptions are considered performance standards or criteria and are entered on the TEF. (Later the evaluator uses this information to decide whether or not the task has been correctly performed.)

In addition to guiding the developer in identifying and describing critical aspects of task performance, the TEF development procedures provide instructions for creating an evaluation scenario and developing a chart for scoring task performance. The evaluation scenario and scoring chart are also entered on the TEF. The evaluator uses the evaluation scenario as instructions for setting up and conducting the evaluation. The scoring chart is used by the evaluator to score the observed task performance.

The individual who develops the TEF will probably not be the same person who uses the TEF to evaluate a task. The person using the TEF development procedures is called the TEF developer. The evaluator is responsible for using the TEF to evaluate task performance. Thus, the developer is actually providing information which will be used later by an evaluator.

A TEF describes for the evaluator:

1. How to set up the evaluation -- an evaluation scenario is described so that all supervisors evaluating the same task will conduct the evaluation in the same manner.

2. What to evaluate -- the task events or steps are designated on the TEF so that all supervisors evaluate the same aspects of task performance. Explicit criteria are provided for determining whether or not an event was successfully completed.

3. How to evaluate -- the scoring chart eliminates evaluator differences in scoring and assigning pass/fail decisions. A separate score is obtained for each evaluation area based upon the number and type of errors which occur during task performance. Explicit criteria for assigning an overall pass/fail decision are also provided.

Examples of Task Evaluation Forms for tasks from three different specialty areas are included in Appendix A.

2.6 Performing the Task Evaluation Form Development Procedures

Throughout the TEF development process, instructions are provided for entering information onto worksheets. At the end of the TEF development procedures, the information from the worksheets is transferred onto Task Evaluation Forms. For each worksheet, the following general information is provided:

1. An overall explanation of the information which will be entered.

2. The purpose of the information from the evaluator's point of view.

3. Examples of completed worksheets.

4. Specific instructions for completing each individual item on the worksheet, including directions for selecting appropriate information and guidelines for entering that information onto the worksheets.

The type of information the developer is required to enter on each worksheet is described below. Sample blank worksheets can be found in Appendix B.

2.6.1 Worksheet 01: Listing the Task Steps

The ueveloper is required to obtain a list of the steps of the task from another source. These steps are referred to throughout the TEF development proce-s.

2.6.2 Worksheet 02: Defining the Task

The first step in Task Evaluation Form development is to define the particular task for which the form is being developed. There are times when the task title does not provide enough information to completely identify the task to be evaluated. Therefore, the following information is provided by the developer:

1. AFSC/Duty Position or Work Center (of the person to be evaluated).

2. Task Title (including the specific version of the task to be evaluated).

3. Task Beginning.

4. Task End.

5. Steps or Events Not Includeu in the Evaluation.

6. Task Information Sources.

2.6.3 Worksheet 03: Evaluation of Time or Speed of Task Performance

The developer applies a set of "criticality questions" to decide whether Time or Speed should be evaluated during the task under consideration. If the developer decides Time or Speed should be evaluated, instructions are provided for identifying exactly when in the task it should be evaluated and the amount of time which is acceptable. The following items are included on Worksheet 03:

1. Critical Segment(s).
2. Starting Point(s).
3. Stopping Point(s).
4. Standard(s).

2.6.4 Worksheet 04: Evaluation of Sequence-Following

The developer applies a set of "criticality questions" to determine whether the order in which the steps of the task are performed should be evaluated. Two types of information are entered on Worksheet 04:

1. Series or sequences of steps in the task which must be performed in order.

2. Single steps in the task which must be performed before other steps.

2.6.5 Worksheet 05: Evaluation of End Products

The developer applies a set of "criticality questions" to determine which outcomes of task performance should be evaluated. Instructions are also provided to aid the developer in identifying criteria for evaluating the end products.

The developer enters the following information onto Worksheet 05:

1. End products.

2. Criteria for evaluating end products.

3. Steps associated with the generation of end products.

2.6.6 Worksheet 06: Evaluation of Safety Procedu es and Regulations

The developer applies a set of "criticality questions" to determine whether adherence to safety procedures and regulations should be evaluated. When specific safety procedures and regulations are identified, the developer follows the instructions for describing the events and for identifying when each event should be evaluated. Thus, the developer enters the following information on Worksheet 06:

1. The safety procedures and regulations which should be evaluated.

2. When to evaluate following safety procedures and regulations.

2.6.7 Worksheet 07: Evaluation of Tools, Equipment, and Materials Use

The developer applies a set of "criticality questions" to determine whether the use of tools, equipment, and materials should be evaluated for the task under consideration. When tools, equipment, or materials are identified, the developer follows the instructions for describing the specific size or type of tools, equipment, and materials which should be used, the correct use of each item, and when in the task the use of each item should be evaluated. Thus, the developer enters the following information on Worksheet 07:

1. Tools, Equipment, and Materials.

2. Size or Type.

3. Correct Use.

4. Steps Associated with Use.

2.6.8 Worksheet 08: Evaluation Scenario

Worksheet 08 is used to describe how the task should be presented to the performer for evaluation purposes. When Worksheet 08 has been completed, an evaluation scenario results. In order to complete the worksheet, the developer describes the:

1. Best Evaluation Method.
2. Preventative Environmental Conditions.
3. Operational Equipment/Task Presentation.
4. Help Permitted.
5. Presentation to Performer of Tools, Equipment, and Materials.
6. Evaluator Equipment.
7. Evaluator Time Estimates.
8. Number of Evaluators.

2.6.9 Additional TEF Development Instructions

Instructions are provided for transferring information from the worksheets onto the Task Evaluation Forms.

When all of the information has been transferred to the TEF, the developer is guided in identifying those events which would result in automatic failure--even when all of the other entries are correctly performed. The developer is instructed to place an asterisk (*) on the TEF next to those events.

2.6.10 Worksheet 09: Scoring Criteria

The final step in TEF development is to assign points to each event on the TEF (points are not assigned to events which result in automatic failure). A formula for deriving the number of points per non-asterisked entry is included on Worksheet 09. Worksheet 09 yields the number of points which should be subtracted for each non-asterisked event that is not successfully performed.

2.7 Presentation of Task Evaluation Form Development Procedures

The TEF development procedures are available for presentation via written handbooks or computer terminals. There are two separate handbooks, titled: Task Evaluation Form Development Procedures for Maintenance and Equipment-Oriented Tasks; and Task Evaluation Form Development Procedures for Non-Equipment Oriented Tasks, AFHRL Technical Papers 85-55 and 85-56, respectively. Both handbooks were developed by Applied Science Associates, Inc. (ASA) under Air Force Contract Number F33615-82-C-0004. Two separate handbooks were prepared to ensure that relevant examples would be provided to a broad range of potential users. The TEF development procedures were adapted for computer delivery by Denver Research Institute in consultation with ASA, under Air Force Contract Number F33615-82-C-0013. Computer delivery of the TEF development procedures allows the user to select relevant examples and help messages. In addition, the computer authoring system provides two levels of support: beginner user and experienced user. This allows experienced TEF developers to select either level of guidance.

The handbooks and the computer authoring system were designed to produce Task Evaluation Forms which are identical in content. Thus, the selection of a presentation method (handbook or computer terminal) will depend upon availability and user preference.

3.0 PROCEDURAL ASSESSMENT

This section describes the efforts made to determine the reliability and validity of the TEF development procedures.

The main objective of the procedural assessment was to examine the outcome of the application of the TEF development procedures to selected tasks in two Air Force specialty areas. Specifically, the reliability and validity of forms resulting from the application of the TEF development procedures were assessed.

A secondary objective of the procedural assessment was to target specific problem areas in the TEF development procedures. Aspects of the TEF content which were relatively low in reliability and/or validity could easily be identified. Once specific problem areas in the procedures were targeted, necessary revisions or enhancements were made. Since the procedural assessment was conducted first in the Aircraft Maintenance career field, it was possible to make some revisions related to targeted problem areas prior to the adaptation of the TEF development procedures for Security Police/Law Enforcement tasks.

For the purposes of this investigation, reliability and validity were defined as follows:

> Reliability referred to the degree of agreement between the content of TEFs developed by two or more Subject-Matter Experts (SMEs) for the same task.
>
> Validity referred to the extent to which the procedures allow SMEs to accurately and completely define the critical evaluation areas, steps, and standards associated with successful task performance.

3.1 Overview

Data related to the procedural reliability and validity were collected in two Air Force specialty areas: Aircraft Maintenance and Security Police/Law Enforcement. Essentially the same procedures for data collection and analysis were used in the two specialty areas.

Subject-Matter Experts applied the Task Evaluation Form development procedures to selected tasks in their career fields. Procedural reliability was assessed by comparing the content of forms prepared for the same task by different developers. Procedural validity was also assessed by examining the content of the TEFs resulting from application of the procedures. Panels of Subject-Matter Experts rated the TEFs with regard to accuracy, completeness, and criticality. Estimates of the validity of the TEF development procedures were obtained by analyzing the rating data. Based upon the results of the procedural analyses, areas for improvement were identified and the necessary changes made.

The data collection, analyses, results, and indicated revisions for both career fields are described in this section.

3.2 Data Collection

The data collection effort involved several phases as follows:

1. Task Selection.
2. Selection of SMEs for TEF Generation.
3. TEF Generation.
4. Panel Rating (data collected from the panel rating were used only in the validity analyses).

Each of these aspects of the data collection effort is described below.

3.2.1 Task Selection

Tasks in each specialty area were chosen to be representative of the types of tasks within the Air Force Specialty Code (AFSC) and to include as much variation in task type and skill level as possible. A total of six tasks were selected to represent the Aircraft Maintenance career field and four tasks were selected from the Security Police/Law Enforcement career field.

3.2.1.1 Aircraft Maintenance Task Selection

Selection of tasks was initially determined by the choice of weapon systems. It was considered necessary to choose weapon systems for which a large amount of task- and training-related documentation existed. On this basis, the Air Force determined that the B-52H and the KC-135A were most appropriate weapon systems for study.

Tasks were selected to represent a cross-section of the types of tasks performed on the weapon systems of interest. Specifically, the chosen tasks represent servicing, inspection, and remove/replace tasks. Tasks representing a range of complexity and skill types were chosen. Other factors considered were frequency of performance and availability of documentation. Based on those factors, three tasks were chosen for each weapon system:

- B-52H

  1. Liquid Oxygen (LOX) Servicing.
  2. Engine Run.
  3. Refuel.

- KC-135A

  1. LOX Servicing.
  2. Preflight.
  3. Starter Cartridge Replacement.

3.2.1.2 Security Police/Law Enforcement Task Selection

Four tasks were selected for TEF generation. Two tasks were selected from the Security Police (SP) field and two tasks were selected from the Law Enforcement (LE) field. An attempt was made to include both tasks which were largely procedural and tasks which were end product or result oriented. The selected tasks are described below:

Building Search (LE). A predetermined scenario was described to the SMEs who participated in the TEF generation. They were instructed to develop TEFs for the following Building Search scenario: "Search Building 20223 for one armed suspect. It is dark and there are no other individuals in the building."

Communications (LE). This task included the preparation for operation and operational check of several different types of radios.

.38 (SP). This task encompassed the loading, disassembly, clearing, and reassembly of the .38.

Handcuffs (SP). This task covered the application and removal of ratchet-type handcuffs with the following scenario: "Suspect is against a wall and does not struggle."

3.2.2 Selection of SMEs for TEF Generation

Subject-Matter Experts (SMEs) from the two specialty areas were selected for participation in the procedural assessment, based upon the following criteria:

1. The SMEs had achieved at least a 5-level in their respective AFSCs.

2. The SMEs were familiar with the tasks under consideration.

3. The SMEs had OJT supervisory experience.

4. The SMEs were available for the time required to generate the TEFs.

Specific details about SME selection in each specialty area are described below.

3.2.2.1 Aircraft Maintenance: SME Selection

Three tasks for each of two weapon systems had been previously selected. Thus, selection of one group of SMEs experienced in B-52H maintenance and another group experienced in KC-135A maintenance was necessary. Available personnel specializing in these weapon systems were found to be concentrated at Carswell Air Force Base, Texas; Ellsworth Air Force Base, South Dakota; and K.I. Sawyer Air Force Base, Michigan. At each base, four SMEs experienced in B-52H maintenance and four SMEs experienced in KC-135A maintenance were selected, for a total of 12 SMEs per weapon system.

3.2.2.2 Security Police/Law Enforcement: SME Selection

Two tasks from the Law Enforcement area and two tasks from the Security Police area had been previously selected. Thus, a group of SMEs representing each area was necessary. Available personnel were found at Kirtland Air Force Base, New Mexico. Thus, all of the data collection effort took place at that location. Six SMEs were selected from the Security Police area and six SMEs were selected from the Law Enforcement area.

3.2.3 TEF Generation

Generation of the Task Evaluation Forms was conducted in a 3-day session at each location. The first day of each session was devoted to SME training. The SMEs were provided with a brief explanation of the project in general, followed by an overview of the TEF development procedures, on the first morning of each session. The afternoon of the first day was dedicated to a practice exercise during which SMEs as a group applied the TEF development procedures to a sample task. Time for questions and discussion was provided throughout the exercise.

The actual TEFs were generated on the second and third days of each session. The SMEs applied the TEF development procedures to one task at a time. SMEs worked independently and were allowed to ask questions concerning only the TEF development procedures. Questions pertaining to the content of the TEFs to be developed were not answered during the TEF generation sessions.

SMEs from the Aircraft Maintenance specialty area were responsible for generating TEFs for each of the three tasks in their respective weapon system. SMEs from Security Police/Law Enforcement areas generated TEFs for the two tasks from their individual areas (i.e., LE or SP).

Twelve replications for each of the Aircraft Maintenance tasks resulted. One of the SMEs from the Law Enforcement area was not able to complete the application of the TEF development procedures. Thus, only five replications of the LE tasks (Communications and Building Search) resulted. Six replications for the .38 and Handcuff tasks were generated.

3.2.4 Panel Rating

Each of the previously generated TEFs was rated by a panel of senior SMEs. The panel members consisted of three SMEs from the appropriate weapon system (KC-135A or B-52H) or area (SP or LE). The panel rating of the forms generated for the Aircraft Maintenance tasks was conducted at Carswell Air Force Base, Texas. Kirtland Air Force Base, New Mexico was selected as the location for the panel rating of the TEFs generated for Security Police and Law Enforcement tasks.

The panel members were selected to meet the following criteria:

1. Each was currently certified to perform the applicable tasks.

2. Each had performed or observed the applicable tasks within the past three months.

3. Each had held substantial responsibility for training or supervising.

4. Each held at least a 7-level in the AFSC under consideration.

5. None had been involved in generating the TEFs.

The panel members were instructed to respond to 15 questions related to the accuracy, completeness, and criticality of the information identified for each evaluation area. A seven-point scale was used. A rating of 1 was considered a low rating for the aspect of validity being rated; 4 was the neutral point; and 7 was the highest rating. Copies of the rating forms are included in Appendix C.

The panel ratings were accomplished in 2-day sessions. The morning of the first day was devoted to orientation and training. The afternoon of the first day and the entire second day were devoted to rating the TEFs.

The panel ratings were assigned independently. The three raters were not permitted to discuss their ratings until the rating was completed for the task under consideration. The actual rating procedures were as follows:

1. Each panel member was given all of the TEFs for one of the tasks in his/her area. The TEFs were prearranged so that all of the panel members received the TEFs in the same order. The panel members were instructed to read through all of the TEFs for the task under consideration before rating any one of the TEFs.

2. The rating sheets for the accuracy dimension were distributed. The panel members were instructed to rate only the accuracy of the information included on the TEFs. It was stressed that accuracy and criticality should not be confused with completeness; a given form could be rated low in completeness and high in criticality and/or accuracy. The panel members were also instructed to indicate on the TEFs any items that were not accurate. When the accuracy ratings were completed, the rating sheets were collected.

3. The rating sheets for the criticality dimension were distributed. Once again, the panel members were instructed to rate only one dimension of validity at a time. The panel members indicated on the TEFs which information was not critical to successful performance of the tasks. When the criticality ratings were completed, the rating sheets were collected.

4. The rating sheets for the completeness dimension were distributed. The panel members were instructed to rate only the completeness of the information on the TEFs. When the completeness of all of the TEFs had been rated, the rating sheets were collected.

5. When the TEFs for a given task had been rated for accuracy, criticality, and completeness, the rating sheets were reviewed for rater discrepancies. A rater discrepancy was said to exist when a given rating differed by 2 or more points from both of the other ratings. A rater whose rating was discrepant was given the opportunity to change his/her rating.

6. All of the rating sheets were recollected.

7. The ratings were discussed as a group.

8. The rating procedures were repeated for the next task of interest.

The panel rating data were used only for the analyses related to the validity of generating the TEFs.

3.3 Data Analysis

During the TEF generation sessions, multiple TEFs were developed for selected tasks in the Aircraft Maintenance career field and in the Security Police/Law Enforcement career field. The content of the TEFs produced for the same tasks was compared to determine the reliability of the TEF development procedures. The TEFs were rated by a panel of senior SMEs. The panel ratings were analyzed to ascertain the validity of the TEF development procedures. The intent of the data analyses was to determine whether the application of the TEF development procedures resulted in forms which were similar in content, when applied by two or more SMEs, and which accurately and completely depicted the critical aspects of task performance.

The specific analyses used to determine the reliability and validity of the TEF development procedures are discussed in detail below.

3.3.1 Reliability Analysis

The reliability of the TEF development procedures is related to the outcome of the application of the procedures to specific tasks. In other words, reliable procedures can be applied by more than one individual (to the same task) with similar results. Thus, the basic approach taken in this study was to examine the result of application of the procedure: the completed TEF. If TEFs produced by two or more SMEs show a high level of agreement in their content, then it can be assumed that the TEF development procedures are reliable.

For the purposes of this study, reliability was defined as the consistency of TEF content.

Estimates of procedural reliability were obtained by dividing the potential information on the TEF into 20 content items and examining the consistency of the information identified for each content item. The consistency of the content items was estimated by calculating mean percentages of agreement for each content item. The content items were then grouped to represent the following aspects of the TEF:

1. Evaluation Area Selection.
2. Time/Speed Content.
3. Sequence-Following Content.
4. End Product Content.
5. Safety Content.
6. Tools, Equipment, and Materials Use Content.
7. Steps Identified.

Separate analyses were performed for the two different career fields (Aircraft Maintenance and Security Police). The two analyses were identical in methodology.

The details of the data analyses are presented below.

3.3.1.1 Mean Percentage of Agreement for Content Items

The potential content of the TEF was divided into 20 content items.

For each content item, the information entered on every form was compared with information included on every other form developed for that task.

A percentage of agreement was calculated for each comparison. Percentages of agreement were calculated as follows:

$$\text{Percentage of Agreement} = \frac{\text{Sum of Matches}}{\text{Sum of Matches + Sum of Mismatches}} \text{ X } 100$$

The content items and their matching criteria are included in Table 1. As shown in the table, certain comparisons were dependent on previous matches. For example, if two TEFs showed no entries for the Time or Speed Evaluation Area, they would score a match for content item 1. However, for subsidiary content items, such as the starting point comparison, neither a match nor a mismatch would be scored. Similarly, when only one of two TEFs contained an entry for an evaluation area, a mismatch would be scored and subsidiary comparisons would not be made.

A mean percentage of agreement for a content item was calculated by averaging the percentages of agreement resulting from all of the comparisons.

When all of the 12 TEFs developed for the Aircraft Maintenance tasks were compared, 66 comparisons per content item resulted. Fifteen comparisons contributed to the mean percentages of agreement for the content items on the .38 and Handcuff TEFs. The mean percentages of agreement for the content items on the Building Search and Communications TEFs were based on 10 comparisons.

3.3.1.2 Content Item Groups

The content items were grouped as follows:

Table 1. Content Items and Matching Criteria

Item	Applicability	Match Criteria
1. Time or Speed	All	Both TEFs (or neither) contain(s) entries for this area.
1a. Steps	All	Both TEFs list same sequence of steps for this measure (within 50% agreement).
1b. Starting Point	Match scored in 1a	Sequence has same starting point (step/event).
1c. Stopping Point	Match scored in 1a	Sequence has same stopping point (step/event).
1d. Standards	Match scored in 1a	Each standard is within 20% of the average of the two standards.
2. Sequence-Following	All	Both TEFs (or neither) contain(s) entries for this area.
2a. Steps	All	Both TEFs list the same sequences of steps for this measure (within 50% agreement).
2b. Starting Point	Match scored in 3a	Sequence has same starting point (step).
2c. Stopping Point	Match scored in 3a	Sequence has same stopping point (step).
3. End Product or Result	All	Both TEFs (or neither) contain(s) entries for this area.
3a. Identified End Products or Results	All	Both TEFs list same item or result.
3b. Steps	Match scored in 2a	Identified end product is associated with same step on both TEFS.

Table 1. Content Items and Matching Criteria (Concluded)

Item	Applicability	Match Criteria
3c. Standards	Match scored in 2a	Same feature is identified or each measured standard is within 20% of the average of the two standards.
4. Safety	All	Both TEFs (or neither) contain(s) entries for this area.
4a. Identified actions or procedures	All	Both TEFs list same action or procedure.
4b. Steps	Match scored in 4a	Identified action or procedure is associated with the same step or steps on both TEFs.
5. Tools, Equipment, and Material Use	All	Both TEFs (or neither) contain(s) entries for this area.
5a. Identified Items	All	Both TEFs list the same tools, equipment, and material.
5b. Steps	Match scored in 5a	Identified item is associated with the same step on both TEFs.
5c. Standards (Correct Use)	Match scored in 5a	Same actions are identified.

Evaluation Area Selection: Iιems 1-5
Time or Speed: Items 1a-d
Sequence-Following: Items 2a-c
End Product: Items 3a-c
Safety: items 4a-b
Tools, Equipment, and Materials Use: Items 5a-c
Ste s: Items 1a, 2a, 3b, 4b, and 5b

The mean percentages of agreement for the applicable content items were averaged to obtain a mean percentage of agreement for each content item group. Results are reported by content item group.

3.3.1.3 Mean Percentages of Agreement for Within and Between Base Comparisons

The TEFs for the Aircraft Maintenance tasks were generated at three different locations. Thus, development location was considered a factor in the analysis of the Aircraft Maintenance data. The following overall mean percentages of agreement were calculated for each task:

1. Within Base.
2. Between Base.
3. Overall.

3.3.2 Validity Analysis

The procedural validity, like the procedural reliability, is related to the outcome of the application of the procedures. In other words, application of the procedures should result in forms which accurately and completely depict the critical aspects of task performance. Once again, the TEFs were used as the basis of the analysis. The validity of the procedures was deduced by investigating the validity of TEFs resulting from procedural application. Specifically, a panel of SMEs rated the TEFs by responding to 15 questions which were related to the accuracy, criticality, and completeness of the information identified for each of the five evaluation areas. A description of the aspects of TEF content rated with each of the questions is contained in Table 2.

For the purposes of this study, validity referred to the accuracy, completeness, and criticality of the information identified for the following aspects of the TEF content:

1. Time/Speed.
2. Sequence-Following.
3. End Product.
4. Safety.
5. Tools, Equipment, and Materials Use.

Table 2. Panel Rating Questions

Question #	Content Rated
1	Accuracy Time/Speed
2	Accuracy Sequence-Following
3	Accuracy End Product
4	Accuracy Safety
5	Accuracy Tools, Equipment, and Materials Use
6	Criticality Time/Speed
7	Criticality Sequence-Following
8	Criticality End Product
9	Criticality Safety
10	Criticality Tools, Equipment, and Materials Use
11	Completeness Time/Speed
12	Completeness Sequence-Following
13	Completeness End Product
14	Completeness Safety
15	Completeness Tools, Equipment, and Materials Use

The three dimensions of validity were defined as follows:

1. Accuracy. The degree to which the information on the TFF is accurate and consistent with Air Force policy.

2. Completeness. The degree to which the information on the TEF is complete.

3. Criticality. The degree to which the information on the TEF is critical to successful task performance.

3.3.2.1 Adjustment for Rater Error

It was mentioned previously that three panel members rated each TEF. Thus, it was necessary to adjust for sources of rater error. A statistical analysis of variance was performed to adjust for rater error. The data were treated as a two-way factorial design (using the ratings as the dependent variable in the ANOVA). Both replications across TEFs and replications across questions were analyzed. The following sources of variance were considered:

1. Between raters (across TEFs).
2. Rater X Form interaction (across TEFs).
3. Between raters (across questions).
4. Rater X Form interaction (across questions).

3.3.2.2 Mean Rating Per Question by Task

Aspects of the validity of the TEFs are represented by different combinations of the panel rating questions. Thus, before specific aspects of validity could be examined, it was necessary to obtain a mean rating per question by task. First, for each question, the three ratings assigned to each form were averaged to obtain a mean rating per question by form. Next, the mean ratings per question by form were averaged across all of the forms developed for a task to obtain a mean rating per question by task.

3.3.2.3 Mean Rating Per Validity Dimension

Estimates of each dimension of validity were obtained by averaging the mean ratings for questions regarding the five evaluation areas by dimension. The rating questions were averaged as follows:

Accuracy	Questions 1-5
Criticality	Questions 6-10
Completeness	Questions 11-15

3.3.2.4 Mean Rating Per Evaluation Area

Individual estimates of the validity of each evaluation area were obtained by averaging the mean ratings for questions regarding each validity dimension by evaluation area. The rating questions were averaged as follows:

Time or Speed - Questions 1, 6, 11
Sequence-Following - Questions 2, 7, 12
End Product - Questions 3, 8, 13
Safety - Questions 4, 9, 14
Tools Equipment, and Materials Use - Questions 5, 10, 15

3.4 Results

The procedural reliability and validity analyses yielded similar patterns of results in the two career fields investigated. As was previously mentioned, revisions were made in the presentation of the TEF development procedures at two points in the study. The first revisions were made after the Aircraft Maintenance assessment, prior to the Security Police/Law Enforcement assessment. Later, revisions were made at the completion of the procedural assessment. In some cases, the interim revisions resulted in improved results for the analyses related to the Security Police/Law Enforcement TEFs. elevant differences in the results found for the two career fields are discussed.

The results of the analyses related to the procedural reliability and validity are presented individually below.

3.4.1 Reliability Results

The results of the following analyses are described:

1. Mean Percentage of Agreement for Evaluation Area Selection.
2. Mean Percentage of Agreement for Time/Speed Content.
3. Mean Percentage of Agreement for Sequence-Following Content.
4. Mean Percentage of Agreement for End Product Content.
5. Mean Percentage of Agreement for Safety Content.
6. Mean Percentage of Agreement for Tools, Equipment, and Materials Use Content.
7. Mean Percentage of Agreement for Steps Identified.
8. Mean Percentages of Agreement for Within and Between Base Comparisons.

3.4.1.1 Mean Percentages of Agreement for Evaluation Area Selection

The mean percentages of agreement for evaluation area selection represent the degree of agreement for identifying evaluation areas which are critical to evaluate for a task. High mean percentages of agreement indicate that the developers identified the same evaluation areas. On the other hand, low mean percentages indicate the developers did not agree regarding the evaluation areas that should be evaluated.

Table 3 shows the mean percentages of agreement for evaluation area selection.

The mean percentages of agreement for evaluation area selection were uniformly high. For the Aircraft Maintenance tasks, the mean percentages of agreement varied from 77 to 91, averaging 85.

With very few exceptions, the developers of the Security Police/Law Enforcement TEFs agreed about the evaluation areas which were critical to each task. As shown in the table, the percentages ranged from 93 to 100, averaging 97.

These mean percentages of agreement indicate that application of the TEF development procedures leads to agreement across developers regarding what aspects of task performance should be evaluated.

3.4.1.2 Mean Percentages of Agreement for Time/Speed Content

These mean percentages of agreement indicate the degree of developer consistency with regard to the identification of information for the Time/Speed evaluation area (content items 1a-d).

In general, the mean percentages of agreement were higher for items associated with Time/Speed than for any other evaluation area. As shown in Table 4, the mean percentages of agreement ranged from 67 to 100, with only three mean percentages falling below 100 for the tasks from both career fields.

This high level of agreement was in part due to the SMEs' decision that this evaluation area was not applicable for seven of the tasks. For tasks where all SMEs agreed that Time or Speed should not be evaluated, the level of agreement was 100 percent.

3.4.1.3 Mean Percentages of Agreement for Sequence-Following Content

These mean percentages of agreement reflect developer consistency regarding the identification of task segments or steps which should be performed sequentially, and the starting and stopping points of those segments (content items 2a-c).

The percentages of agreement are shown in Table 5.

Table 3. Mean Percentages of Agreement--Evaluation Area Selection

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	89
B-52H Refuel	83
B-52H LOX	91
KC-135A Starter Cartridge Replacement	89
KC-135A LOX	77
KC-135A Preflight	80
Overall	85
	Security Police/Law Enforcement Tasks
Building Search	100
Communications	100
.38	93
Handcuffs	93
Overall	97

Table 4. Mean Percentages of Agreement--Time or Speed Content

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	100
B-52H Refuel	100
B-52H LOX	82
KC-135A Starter Cartridge Replacement	100
KC-135A LOX	92
KC-135A Preflight	100
Overall	96
	Security Police/Law Enforcement Tasks
Building Search	100
Communications	100
.38	67
Handcuffs	100
Overall	92

Table 5. Mean Percentages of Agreement--Sequence-Following Content

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	61
B-52H Refuel	94
B-52H LOX	79
KC-135A Starter Cartridge Replacement	77
KC-135A LOX	80
KC-135A Preflight	82
Overall	79
	Security Police/Law Enforcement Tasks
Building Search	80
Communications	91
.38	75
Handcuffs	100
Overall	86

For the Aircraft Maintenance tasks, the percentages of agreement averaged 79. For three of the tasks, most SMEs agreed that the entire task should be performed sequentially (B-52H Refuel, B-52H LOX, and KC-135A LOX). However, only a small number of SMEs identified sequences for the remaining three tasks. In some cases, the SMEs agreed that the steps should be performed sequentially, but disagreed regarding the stopping point. These discrepancies somewhat lowered otherwise high percentages of agreement.

In general, the Security Police/Law Enforcement developers agreed regarding the sequences to be evaluated and their starting and stopping points. The mean percentages of agreement ranged from 75 to 100, averaging 86. For the tasks which had less than 100 percent agreement, all of the discrepancies were minor (e.g., a starting or stopping point off by one step).

3.4.1.4 Mean Percentages of Agreement for End Product Content

These mean percentages of agreement depict the degree of developer consistency regarding the end products identified, criteria for evaluating those end products, and the steps at which the end products should be evaluated (content items 3a-c).

The mean percentages of agreement for end product content are shown in Table 6. These figures ranged from 69 to 82 for the Aircraft Maintenance tasks and 62 to 82 for the Security Police/Law Enforcement tasks.

Figures for this evaluation area were lowered by the disparate number of end products identified on different forms. For example, a TEF with few end products listed, when compared with a TEF showing many end products, produced a high number of mismatches, even if the listed end products were matched. For cases where a match was found, subsidiary content items showed consistently high levels of agreement. In other words, when two SMEs identified the same end product, they also identified the same associated steps and criteria. In general, levels of agreement for the end product content remained in the acceptable range.

3.4.1.5 Mean Percentages of Agreement for Safety Content

The mean percentages of agreement for safety content reflect the degree of agreement regarding the safety procedures and regulations which were identified as critical to successful task performance and the steps at which they should be evaluated (content items 4a and 4b).

As indicated in Table 7, Safety had lower percentages of agreement than any other evaluation area. Percentages of agreement ranged from 52 to 68 for Aircraft Maintenance tasks. The Security Police/Law Enforcement tasks obtained mean percentages of agreement ranging from 47 to 63.

Table 6. Mean Percentages of Agreement--End Product Content

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	79
B-52H Refuel	82
B-52H LOX	72
KC-135A Starter Cartridge Replacement	76
KC-135A LOX	69
KC-135A Preflight	76
Overall	76
	Security Police/Law Enforcement Tasks
Building Search	82
Communications	79
.38	72
Handcuffs	62
Overall	74

Table 7. Mean Percentages of Agreement--Safety Content

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	68
B-52H Refuel	57
B-52H LOX	52
KC-135A Starter Cartridge Replacement	54
KC-135A LOX	58
KC-135A Preflight	60
Overall	58
	Serurity Police/Law Enforcement Tasks
Building Search	63
Communications	47
.38	47
Handcuffs	51
Overall	52

Once again, the low figures were primarily due to variation in the number of safety events identified. When a TEF with many events identified was compared with a TEF with few entries, a high number of mismatches inevitably occurred. However, when the developers agreed about the inclusion of a safety event, they also obtained high levels of agreement regarding when that safety event should be evaluated.

3.4.1.6 Mean Percentages of Agreement for Tools, Equipment, and Materials Use Content

These mean percentages of agreement depict the degree of evaluator agreement regarding entries for the evaluation of the use of tools, equipment, and materials. Specifically, developer consistency for identifying the items for evaluation, the steps at which the items should be evaluated, and the correct use of the identified items are reflected in these figures.

The mean percentages of agreement obtained for the Security Police/Law Enforcement tasks were substantially higher than those obtained for the Aircraft Maintenance tasks. In fact, the lowest single task percentage of agreement was obtained for the B-52H Refuel task (31). Several factors affected developer agreement for this evaluation area. First, the developers were confused about what items qualified as tools, equipment, or materials. Second, the developers were not certain about the type of information to include (i.e., the the correct size/type of item, the correct use, or both). Finally, when the use of an item was already included in another evaluation area, the developers were uncertain about including that item under the tools, equipment, and materials use evaluation area. These problems drastically reduced the reliability figures for the forms developed for the Aircraft Maintenance tasks.

Revisions were made in the instructions related to this particular evaluation area prior to the generation of the forms for the Security Police/Law Enforcement tasks. The resulting increase in the level of agreement for the information entered on the Security Police/Law Enforcement TEFs is illustrated in Table 8. The percentages of agreement ranged from 71 to 76 for the Security Police/Law Enforcement tasks, compared to 31 to 78 for the Aircraft Maintenance tasks.

3.4.1.7 Mean Percentage of Agreement for Steps Identified

The mean percentages of agreement for steps identified are the averages of the content items (1a, 2a, 3b, 4b, and 5b) related to the steps at which events should be evaluated. High levels of agreement indicate that, for all evaluation areas, identification of critical steps can be reliably accomplished. On the other hand, low levels of agreement indicate that once an event was identified for inclusion in

Table 8. Mean Percentages of Agreement--Tools, Equipment, and Materials Use

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	59
B-52H Refuel	31
B-52H LOX	61
KC-135A Starter Cartridge Replacement	78
KC-135A LOX	65
KC-135A Preflight	76
Overall	62
	Security Police/Law Enforcement Tasks
Building Search	76
Communications	73
.38	71
Handcuffs	71
Overall	73

an evaluation area, developer disagreement occurred regarding when in the task it should be evaluated.

Table 9 shows the percentages of agreement for step-related content items, averaged across evaluation areas.

The mean percentages of agreement were somewhat higher for the Aircraft Maintenance tasks, averaging 85, compared to 80 for Security Police/Law Enforcement tasks. However, all of the figures demonstrate acceptable levels of agreement.

3.4.1.8 Mean Percentages of Agreement Within Base and Between Bases

The Aircraft Maintenance analysis incorporated comparisons of the agreement levels within bases and between bases. As Table 10 shows, mean percentages of agreement were generally somewhat higher for within-base comparisons compared to between-base comparisons. However, the between-base figures were high enough to demonstrate an adequate level of agreement. In practical terms, a TEF developed at one base will be similar to one developed at another base.

3.4.2 Validity Results

The results of the following analyses are described:

1. Adjustment for Rater Error.
2. Mean Ratings per Validity Dimension.
3. Mean Ratings per Evaluation Area.

3.4.2.1 Adjustment for Rater Error

The analyses of variance did not reveal sources of rater error for the data related to either career field. Thus, adjustments for rater error were not required.

3.4.2.2 Mean Ratings by Validity Dimension

Accuracy. The raters were asked to rate, on a scale of 1 to 7, the content of the TEFs with regard to accuracy. A rating of 7 indicated that the information on the TEF was accurate and consistent with Air Force policy. On the other hand, a rating of 1 meant that the information was not consistent with Air Force policy.

As shown in Table 11, the mean ratings for accuracy did not fall below 6.2 for any task. These high mean ratings for accuracy indicate that the TEF development procedures result in the identification of information which is accurate and consistent with Air Force policy.

Table 9. Mean Percentages of Agreement--Steps Identified

Task	Percentage
	Aircraft Maintenance Tasks
B-52H Engine Run	85
B-52H Refuel	77
B-52H LOX	92
KC-135A Starter Cartridge Replacement	81
KC-135A LOX	87
KC-135A Preflight	89
Overall	85
	Security Police/Law Enforcement Tasks
Building Search	96
Communications	75
.38	71
Handcuffs	79
Overall	80

Table 10. Mean Percentages of Agreement--Within Base and Between Bases

Content Item Group	Task	Percentage Within Base	 Between Bases	 Overall
Evaluation Area Selection				
	B-52H Engine Run	96	87	89
	B-52H Refuel	89	81	83
	B-52H LOX	94	90	91
	KC-135A Starter Cartridge Replacement	93	88	90
	KC-135A LOX	79	76	77
	KC-135A Preflight	[illegible]	79	80
Time or Speed				
	B-52H Engine Run	100	100	100
	B-52H Refuel	100	100	100
	B-52H LOX	83	83	82
	KC-135A Starter Cartridge Replacement	100	100	100
	KC-135A LOX	92	93	92
	KC-135A Preflight	100	100	100
Sequence-Following				
	B-52H Engine Run	60	61	61
	B-52H Refuel	94	94	94
	B-52H LOX	80	80	80
	KC-135A Starter Cartridge Replacement	83	75	77
	KC-135A LOX	81	80	80
	KC-135A Preflight	41	82	82
End Products				
	B-52H Engine Run	82	77	79
	B-52H Refuel	84	81	82
	B-52H LOX	75	72	73
	KC-135A Starter Cartridge Replacement	80	75	77
	KC-135A LOX	70	68	69
	KC-135A Preflight	78	75	76

Table 10. Mean Percentages of Agreement--Within Base and Between Bases (Concluded)

Content Item Group	Task	Percentage		
		Within Base	Between Bases	Overall
Safety				
	B-52H Engine Run	70	68	68
	B-52H Refuel	69	63	57
	B-52H LOX	61	48	52
	KC-135A Starter Cartridge Replacement	60	51	54
	KC-135A LOX	63	57	58
	KC-135A Preflight	61	59	60
Tools, Equipment, and Materials Use				
	B-52H Engine Run	54	63	59
	B-52H Refuel	39	50	31
	B-52H LOX	69	58	61
	KC-135A Starter Cartridge Replacement	83	77	78
	KC-135A LOX	66	66	65
	KC-135A Preflight	81	74	76
Steps Identified				
	B-52H Engine Run	85	85	85
	B-52H Refuel	77	94	77
	3-52H LOX	95	92	92
	KC-135A Starter Cartridge Replacement	85	79	81
	KC-135A LOX	85	88	87
	KC-135A Preflight	88	89	90

Table 11. Mean Ratings by Validity Dimension

	Validity Dimensions		
Aircraft Maintenance Tasks	Accuracy	Completeness	Criticality
B-52H Engine Run	6.3	4.5	6.2
B-52H Refuel	6.2	4.8	6.2
B-52H LOX	6.4	4.9	6.4
KC-135A Starter Cartridge Replacement	6.4	4.9	6.6
KC-135A LOX	6.6	4.8	6.5
KC-135A Preflight	6.2	4.2	6.5
Overall:	6.4	4.7	6.4
Security Police/Law Enforcement Tasks			
Building Search	6.2	7.0	6.7
Communications	6.5	5.7	6.7
.38	6.6	6.4	7.0
Handcuffs	6.9	5.9	6.8
Overall:	6.6	6.2	6.9

Table 12. Mean Ratings by Evaluation Area

	Evaluation Areas				
Aircraft Maintenance Tasks	Time/ Speed	Sequence-Following	End Product/ Result	Safety	Tools, Equipment, and Materials
B-52H Engine Run	7.0	5.4	6.0	5.4	3.5
B-52H Refuel	7.0	6.7	5.7	4.1	2.0
B-52H LOX	6.8	6.6	5.6	5.7	4.4
KC-135A Starter Cartridge Replacement	6.3	4.3	5.9	5.8	6.0
KC-135A LOX	5.7	6.4	5.4	6.0	4.5
KC-135A Preflight	6.8	4.1	6.6	5.3	3.7
Overall:	6.6	5.6	5.9	5.4	4.0
Security Police/Law Enforcement Tasks					
Building Search	6.6	7.0	7.0	6.8	6.4
Communications	6.7	6.2	6.4	5.8	6.5
.38	7.0	6.8	6.6	6.4	6.6
Handcuffs	7.0	7.0	6.4	6.0	6.3
Overall:	6.8	6.7	6.6	6.2	6.5

Completeness. The raters assigned a rating of 1 to 7 to the completeness of the information entered on the TEFs. A rating of 7 represented a high level of completeness and a rating of 1 indicated that the TEFs were substantially incomplete.

The completeness ratings for the Aircraft Maintenance tasks were considerably lower than the ratings for the other two dimensions of validity. The mean ratings ranged from 4.2 to 4.9. Although all of the mean ratings fell in the valid range (above the neutral point of 4), completeness of the TEFs was targeted as a problem area. Revisions aimed at increasing the completeness of the TEFs were made in the presentation of the TEF development procedures prior to the generation of the forms for the Security Police/Law Enforcement tasks. Apparently, those revisions remedied the previous problems with completeness. As Table 11 shows, the mean completeness ratings for the Security Police/Law Enforcement forms were substantially improved and more consistent with the mean ratings for accuracy and criticality.

Criticality. The raters assigned a rating of 1 to 7 to the criticality of the information included on the TEFs. A rating of 7 meant that the information was critical to successful task performance. On the other hand, a rating of 1 indicated that the information was not critical to successful task performance and thus should not be included on the TEFs.

The mean criticality ratings were uniformly high, ranging from 6.2 to 7.0. The criticality ratings indicate that application of the TEF development procedures results in the identification of only those events which are critical to successful task performance.

3.4.2.3 Mean Ratings by Evaluation Area

The mean ratings for accuracy, completeness, and criticality were averaged to obtain mean ratings by evaluation area for each task. Table 12 shows the mean ratings for each evaluation area.

The mean ratings for the Time/Speed and End Product evaluation areas were consistently high. Relatively lower ratings were found for the other evaluation areas for the Aircraft Maintenance tasks. However, all of the relatively lower mean ratings can be attributed to the averaging of the low completeness ratings with the higher ratings for the other two dimensions. In spite of the negative contributions of the low completeness ratings, most of the ratings remained in the valid range (above the neutral point of 4). The only ratings which fell below 4.0 were ratings of the Tools, Equipment, and Materials Use content for the Aircraft Maintenance tasks. The problems with this evaluation area and with completeness, in general, were addressed through revisions in the presentation of the TEF development procedures prior to the generation of the TEFs for the Security Police/Law Enforcement tasks. The success of those revisions is reflected in the developer's knowledge. Secondly, automation should alleviate some problems associated with incomplete application of the procedures. The TEF development procedures automation effort was discussed in Section 2.0.

Ratings of the accuracy and criticality of TEFs were consistently high, indicating that application of the TEF development procedures results in the identification of information which is consistent with Air Force policy and critical to successful task performance. Ratings of the completeness dimension were relatively low for the Airc aft Maintenance tasks. This problem area was addressed through revision in the presentation of the TEF development procedures. The completeness ratings for the Security Police/Law Enforcement task showed substantial improvements.

In general, the validity analyses were positive, demonstrating that the TEF development procedures can indeed be used to generate valid forms. The primary targeted problem area (low completeness ratings) was improved substantially through revisions to the handbooks.

Taken together, the results of the procedural reliability and validity analyses indicate that when the TEF development procedures are applied to specific tasks by SMEs, the resulting forms contain information which is both reliable and valid.

substantially improved mean ratings by evaluation area obtained for the content of the Security Police/Law Enforcement forms. Only one of the mean ratings fell below 6.0 (5.8 for the Safety evaluation area of the Communications task).

In summary, the application of the TEF development procedures resulted in the identification of valid information in most cases. When interim revisions were made in the presentation of the TEF development procedures, accurate, complete, and critical information was identified for all evaluation areas.

3.5 Summary and Conclusions

Overall, the data from the procedural reliability analyses, in both career fields, demonstrated a high level of reliability in the TEF development procedures. The percentages of agreement from the Security Police analysis showed some improvement over the Aircraft Maintenance analysis, when considered overall. This improvement demonstrates the applicability of the procedures to the different task areas, as well as suggesting that the interim revisions to the handbook were effective. It should be noted that the smaller sample size of the Security Police analysis made it more vulnerable to error. In particular, discrepancies caused by a single SME's opinion had a greater impact in the Security Police analysis than in the Aircraft Maintenance analysis. In this light, the overall improvement is even more noteworthy.

In general, the highest area of reliability was in the identification of evaluation areas which are applicable to task performance. The identification of specific events in an evaluation area obtained relatively less agreement. In particular, low levels of agreement were associated with the identification of specific safety procedures and regulations to be evaluated. However, once entries were identified, high levels of agreement for the steps and standards associated with those items resulted.

Based upon the aforementioned results, changes were made in the presentation of the TEF development procedures which should result in substantially improved reliability of TEFs generated in the future. Specifically, instructions for identifying task end products and safety procedures and regulations for evaluation purposes have been added. In general, the clarity, specificity of the instructions, and the ease of procedural application was enhanced.

Obviously, actual reliability will be dependent to a certain extent on the procedures' utilization. Specific factors ha e been pinpointed that will optimize reliability. First, it may be appropriate for TEFs to be developed by SME teams rather than individuals. The consensus opinion of a well-motivated group of developers should provide a more reliable source than any individual

4.0 OPERATIONAL ASSESSMENT

In the previous section, the reliability and validity of the development procedures were discussed. This section details the activities related to the operational assessment of the TEFs.

The objective of the operational assessment was to examine the use of the TEFs to evaluate OJT tasks performed in two Air Force specialty areas. Specifically, the reliability and validity of the results of TEF evaluations were assessed. For the purposes of the operational assessment, reliability and validity were defined as:

1. Reliability - the consistency of error detection, end scores, and pass/fail assignments within each evaluator group.

2. Validity - the degree to which evaluations result in error detection and to which the pass/fail assignments and end scores discriminate between performance levels.

4.1 Overview

The use of the TEFs was compared with the use of traditional methods of evaluating OJT task performance. Two groups of evaluators observed a high and low performance level of two tasks in their specialty areas. One group of evaluators used the TEFs to evaluate task performance and another group used traditional methods (checklists) for evaluating task performance. The results of the two types of evaluation were compared to determine the relative reliability and validity of each evaluation method.

4.2 Data Collection

The following activities were conducted in preparation for the operational assessment:

1. Task Selection.
2. TEF Generation.
3. Error Selection.
4. SME Selection.
5. Performer Instruction.
6. Evaluator Instruction.
7. Task Performances and Evaluation.

4.2.1 Task Selection

Two tasks were selected from the Security Police/Law Enforcement area and two tasks were selected from the Aircraft Maintenance area. The criteria for task selection were:

1. The tasks could be observed by multiple evaluators.
2. The tasks could be performed within a reasonable time frame to allow the repetitions required by the research design.
3. The tasks contained "critical" elements for a variety of evaluation areas.
4. The tasks required a minimum of "rigging" or preparation prior to each performance.
5. The tasks allowed the inclusion of errors in task performance without threatening the safety of the performer, evaluator, or equipment.

Using these criteria as guidelines, the following tasks were selected to be performed during the operational assessment:

Aircraft Maintenance

Preflight KC-135A
Refuel B-52H

Security Police/Law Enforcement

Application and Removal of Ratchet-Type Handcuffs
Issue and Turn-In of M-16

4.2.2 TEF Generation

4.2.2.1 Aircraft Maintenance

Task Evaluation Forms for the Refuel and Preflight tasks were generated during the procedural assessment. The forms which received the highest ratings for reliability and validity were selected for use in the operational assessment. A panel of SMEs was asked to review these forms and make additions to ensure their completeness.

4.2.2.2 Security Police/Law Enforcement

Task Evaluation Forms had been generated during the procedural assessment for one of the two selected tasks. The TEF for the Handcuff task which received the highest ratings was used in the operational assessment. A panel of SMEs reviewed the form and made additions to ensure its completeness.

Since a TEF for the M-16 task did not exist, a team of SMEs was asked to generate one. A second group of SMEs reviewed the resulting TEF and made additions to ensure its completeness.

4.2.3 Error Selection

A panel of Subject-Matter Experts from each specialty area was asked to compile a list of errors which typically occur during OJT performance of the selected tasks. Errors from this list were then selected for insertion into the performances durinq the operational assessment.

Two types of errors were included in the task performances. The error types are described below.

Type 1 error. Errors which would result in points subtracted from the final score but would not result in automatic failure.

Type 2 error. Errors which would result in automatic failure even if everything else had been performed correctly.

Errors which would directly endanger the performer, the evaluators, or equipment were not selected. However, some errors which would normally cause cessation of the evaluation were selected so that detection of all levels of errors could be assessed. For performances involving this type of error (Type 2), evaluators were asked to record the errors and to continue with the evaluation.

Two separate performances of each task were staged. The first performance was a "good" performance containing only Type 1 errors. The second performance was a "bad" performance which could contain both Type 1 and Type 2 errors.

4.2.4 SME Selection

Subject-Matter Experts (SMEs) from the two specialty areas were selected for participation in the operational assessment based upon the following criteria:

1. The SMEs had achieved at least a 5-level in their respective AFSCs.

2. The SMEs were familiar with the task(s) to be evaluated.

3. The SMEs had OJT supervisory experience.

4. The SMEs were available for the time required by the research design.

In order to make comparisons between the two evaluator groups, a minimum of eight evaluators were required for each group. Thus, at a minimum, 16 SMEs were necessary to evaluate each task. In addition, SMEs were required to stage the task performances.

The operational assessment methodology was repeated in both specialty areas. Specific details about SME selection in each specialty area are provided below.

4.2.4.1 SME Selection--Aircraft Maintenance

One task from each of two weapon systems had been previously selected. Thus, one group of SMEs experienced in B-52H maintenance and another group experienced in KC-135A maintenance were necessary.

For each task, a minimum of eight evaluators were required per evaluator group. Additional SMEs were necessary to enact the staged task performances. Thus, at a minimum, 17 SMEs were necessary for each task, yielding a total of 34 SMEs.

This number of SMEs was not available at Carswell Air Force Base, Texas, the original location selected for the operational assessment. A second location at Dyess Air Force Base, Texas, was added.

At each of the two locations, 18 SMEs participated in the project. Nine SMEs from each base were experienced in B-52H maintenance and the other nine were experienced in KC-135A maintenance. SME participation for each task was broken down as follows:

1 Performer
4 Checklist Evaluators
4 TEF Evaluators

At each location, the same performer staged both the "good" and "bad" task performances. The evaluators observed both performances of their respective tasks. When the evaluators from both locations were combined, a total of eight evaluators per evaluator group resulted.

4.2.4.2 SME Selection--Security Police/Law Enforcement

Kirtland Air Force Base, New Mexico, was selected as the location for the operational assessment. It was possible to find SMEs at that location who were experienced in both of the tasks which had been previously selected.

SME participation from the Security Police/Law Enforcement area was broken down as follows:

2 Performers
10 Checklist Evaluators
10 TEF Evaluators

Two performers were necessary to enact the task performances. The same performers staged all of the task performances (the "good" and "bad" performances of both tasks). Also, the same evaluators observed all of the task performances. Thus, a total of 10 SMEs per evaluator group resulted.

4.2.5 Performer Instruction

The performers received a brief orientation to the project as well as specific instructions regarding the task performances. During the orientation, the performers were told that the purpose of the operational assessment was to assess hypothetical task performances which would not reflect their own ability to perform the task. The specific instructions regarding the task performance for each level ("good" and "bad") included:

1. Information regarding the steps of the task to be performed.

2. A description of the errors to be inserted in each version of the task.

3. A discussion of the importance of standardizing performance across evaluator groups.

4.2.6 TEF Evaluator Group Instruction

The TEF evaluator group used the Task Evaluation Forms to evaluate task performance. This group received training in the use of the TEFs, including:

1. An opportunity to become familiar with the specific TEFs used for the operational assessment.

2. Instructions for using the forms to observe the task performances.

3. Instructions for scoring the forms.

4. An opportunity to review the Evaluator Handbook (which included the scoring and application procedures).

5. Information regarding what was required during the operational assessment.

   a. Scheduling.
   b. Evaluating independently.
   c. Deriving a score.

4.2.7 Checklist Evaluator Group Instruction

The checklist evaluator group used checklists to evaluate task performance. They were not provided with training in the use or scoring of the checklists. This training was omitted based upon the rationale that the purpose of the study was to compare the TEF and checklist methods of evaluation. As such, it was preferred that both methods be applied as realistically as possible. The current practice of using the checklist without scoring or application instruction was maintained for the purposes of the study.

SMEs selected for the checklist evaluator group were required to have previously performed evaluations with the checklists. Additional training included:

1. An opportunity to become familiar with the specific checklists used in the operational assessment.

2. Instructions for completing the checklist during the task performances.

3. Information regarding what was required during the operational assessment:

   a. Scheduling.

   b. Evaluating independently.

   c. Arriving at an end score (i.e., rating the performance on a scale of 1 to 10 in order to facilitate comparison between the scores obtained by TEF and checklist evaluator groups).

4.2.8 Specialty Area Differences

The operational assessment methodology was identical in the Security Police/Law Enforcement and Aircraft Maintenance career fields, with the exception of the changes mentioned below:

1. The operational assessment in the Aircraft Maintenance career field was conducted at two locations, with one-half of each evaluator group derived from each base. All of the SMEs from the Security Police/Law Enforcement specialty area came from Kirtland Air Force Base, New Mexico.

2. There was a total of eight evaluators per group for the Aircraft Maintenance tasks and 10 for the Security Police/Law Enforcement tasks.

3. The same performers and evaluators participated in the assessment of the Handcuff and M-16 tasks. On the other hand, the performers and evaluators from the Aircraft Maintenance specialty area participated in the assessment of either the Preflight task or the Refuel task. A separate group of performers and evaluators was necessary for each weapon system.

4.2.9 Task Performances and Evaluation

Once training of the performers and evaluators had been completed, the task performances and evaluations commenced. For each task at each base, the TEF and checklist groups evaluated task performances separately. This separation was due to the different observation and completion requirements of the checklists and the TEFs. To ensure consistency between performances for the two evaluator groups:

1. The same SMEs performed the task for both evaluator groups.

2. The SMEs were trained to perform only those errors which were predesignated.

3. The SMEs were reminded that the two evaluator group performances should be identical.

4. Contractor personnel observed all task performances.

Evaluators were asked to conduct the evaluation as if they were conducting an actual OJT performance assessment. The only exceptions were that there were multiple evaluators (evaluating independently), and that the checklist group scored the performance on a scale of 1 to 10 after the evaluation had been completed.

4.3 Data Analysis

During the task performances of the operational assessment, data were collected on the results of evaluations using checklists as well as TEFs. The checklists represented the current method of OJT evaluation. The outcomes of the two types of evaluations were compared in a series of analyses. Checklist data were used for comparison purposes only, and were not meant to represent a criterion for OJT evaluation instruments. Data were collected from both methods of evaluation regarding errors detected, end score derived, and overall pass/fail assignment. The intent of the data analysis was to determine whether the use of the TEF resulted in an improvement over the current evaluation method with regard to the reliability and/or validity of the evaluation results.

The definitions and data analysis used to determine the relative reliability and validity of the two evaluation methods are described in detail in this section.

4.3.1 Reliability Analysis

The reliability of the use of an OJT evaluation instrument represents the ability of the instrument to provide consistent evaluation results. In other words, two or more evaluators observing the same performance should derive the same evaluation results, including errors detected, end scores derived, and the pass/fail decisions assigned.

For the purposes of this study, reliability was defined as the consistency of error detection, end scores derived, and pass/fail assignments between evaluators in the same evaluator group.

Estimates of the reliability of the two evaluation methods were obtained by considering the following:

1. The consistency of error detection within evaluator groups which was estimated by calculating mean percentages of agreement.

2. The relative consistency within evaluator groups of end score assignment which was determined by calculating variance ratios for the amount of variance in the end scores assigned by each evaluator group.

3. The relative consistency of pass/fail assignments which was assessed by computing the percentage of pass/fail assignments made by each evaluator group by task performance.

The individual data analyses in the two specialty areas were identical. These analyses are described in detail below.

4.3.1.1 Mean Percentages of Agreement

The consistency of error detection among evaluators assigned to an evaluator group was determined by calculating a mean percentage of agreement for error detection for each task performance.

The formula used to calculate the mean percentages of agreement was identical to the formula utilized in the procedural reliability assessment. The formula is shown below.

$$\text{Percentage of Agreement} = \frac{\text{\# matches}}{\text{\# mismatches} + \text{\# matches}} \times 100$$

The errors detected by each evaluator were compared to the errors detected by each of the other evaluators in the group. A match was scored when a pair of evaluators detected the same error or when a pair of evaluators missed the same error. When only one of the pair detected an error, a mismatch was scored.

When the errors detected by each of eight evaluators who observed the Aircraft Maintenance task performances were compared, 28 percentages of agreement resulted. Comparison of the errors detected by each of the 10 Security Police/Law Enforcement evaluators per group yielded 45 percentages of agreement.

The percentages of agreement were averaged to calculate the mean percentages of agreement by task for an evaluator group. A mean percentage of agreement was obtained for each of the task performances, yielding a total of eight mean percentages of agreement per evaluator group.

4.3.1.2 Variance Ratios

The consistency of end score assignments within evaluator groups was compared by examining the amount of variance in the end scores assigned by the two groups. The amount of variance in end scores obtained by the TEF group was compared to the variance in end scores obtained by the checklist group by calculating variance ratios. The ratios were calculated by placing the largest variance (regardless of evaluator group) in the ratio's numerator. A variance ratio was calculated for every task performance, yielding a total of eight variance ratios (four for performances of Aircraft Maintenance tasks and four for performances of the Security Police/Law Enforcement tasks).

Significance levels were determined for each of the eight variance ratios.

4.3.1.3 Percentage of Pass/Fail Assignments

The consistency of pass/fail assignments within evaluator groups was compared by examining the percentages of passes and fails assigned tc each performance, by the two groups.

4.3.2 Validity Analyses

The validity of the use of an OJT evaluation instrument represents tne ability of that instrument to provide evaluation results which accurately reflect observed performance. In other words, use of the instrument should result in detection of performance errors. In addition, end scores and pass/fail assignments should reflect the level of performance observed; i.e., higher end scores and more pass decisions should be assigned to the "better" performances, with lower end scores and more fail decisions assigned to the "worse" performances.

For the purposes of this study, validity was defined as the accuracy of error detection and the discrimination between performance levels in end scores and pass/fail assignments.

Estimates of the validity of the two evaluation methods were obtained by considering the following:

1. The ability of the two evaluator groups to detect errors in performance which was compared by _t_-tests between the mean number of errors detected by each evaluator group for each task performance.

2. The ability of two evaluator groups to discriminate between levels of performance with regard to end score assignment which was determined by _t_-tests between the mean end scores assigned to the high and low task performances for each evaluator group.

3. The ability of each evaluator group to discriminate between levels of performance with regard to pass/fail decisions which was examined by calculating chi-squares. The chi-squares tested the null hypothesis of no discrimination between levels of performance in pass/fail decisions.

4. The ability of the evaluator groups to assign end scores which reflect the number of errors detected in task performance which was determined by calculating correlations between end scores derived and the number of errors detected in task performance.

Once again, the data analyses for the two specialty areas (Aircraft Maintenance and Security Police/Law Enforcement) were identical. Details of the analyses are provided below.

4.3.2.1 T-Tests for the Difference Between the Mean Number of Errors Detected

The ability of the two evaluator groups to detect performance errors was compared by performirg _t_-tests comparing the mean numbers of errors detected.

The _t_'s were calculated for each task performance, yielding a total of eight _t_'s (four for the task performances in each specialty area). Significance levels were determined for each _t_-test.

4.3.2.2 T-Tests for the Difference Between End Scores for High and Low Performances

The ability of evaluators to discriminate between high and low performance levels with regard to end scores was assessed by performing _t_-tests for the difference between the end scores derived for the "good" and "bad" task performances. For each evaluator group, one _t_-test was performed for each task. Thus, four _t_'s resulted for each evaluator group (two _t_'s represented Aircraft Maintenance tasks and two _t_'s represented Security Police/Law Enforcement tasks). Significance levels were determined for each _t_-test.

4.3.2.3 Chi-Square Discrimination Between Performances in Pass/Fail Assignments

The ability of the two evaluator groups to assign pass/fail decisions which discriminated between performance levels was assessed by calculating chi-squares. One chi-square was performed for each task, resulting in four chi-squares for each evaluator group. The chi-squares tested the null hypothesis of no discrimination between levels of performance in pass/fail decisions. Significance levels were determined for each chi-square.

4.3.2.4 Correlations - Numbers of Errors and End Score

The degree of sensitivity of the end scores to changes in the number of errors detected was determined by computing correlations between the end scores derived and the number of errors detected in task performance.

Pearson product-moment correlations were calculated by task performance. Eight correlations were computed for the checklist evaluator group. Four of the eight correlations represented Aircraft

Maintenance task performances and the other four represented Security Police/Law Enforcement task performances. It was not possible to complete the correlations for three of the "bad" performances observed by the TEF group because all of the TEF evaluators assigned end scores of 0 to those performances. Thus, only five correlations were computed for the TEF results. Significance levels were determined for all of the correlations.

4.4 Results

The results of each of the analyses will be described separately. The data analyses revealed similar results for the two specialty areas. Thus, the discussions of the results for the Aircraft Maintenance tasks and Security Police tasks will be combined.

4.4.1 Reliability Results

The results of the following analyses are included:

1. Mean Percentages of Agreement.
2. Variance Ratios.
3. Percentage of Pass/Fail Assignments.

4.4.1.1 Mean Percentages of Agreement

The mean percentages of agreement depict agreement within evaluator groups. High mean percentages of agreement indicate that the evaluators detected the same errors. On the other hand, low mean percentages of agreement indicate that the evaluators did not detect the same errors. A word of caution is necessary before interpreting the mean percentages of agreement. Due to the nature of the formula (division of the number of matches plus mismatches by the number of matches and the low number of matches), mean percentages of agreement were easily affected when only one or two of the evaluators missed or detected an error. For example, in the Aircraft Maintenance analyses, only 28 comparisons contributed to each mean percentage of agreement. Therefore, when one of the evaluators in this specialty area consistently scored a mismatch with the other evaluators, seven of the 28 (25%) percentages of agreement were affected. In addition, the low possible number of matches adversely affected the "good" performances of all four tasks, which had only two or three possible matches.

The mean percentages of agreement should be used to compare the agreement within the TEF group to agreement within the checklist group rather than to estimate the reliability of either group's results. In addition, the mean percentages of agreement reflect only agreement or consistency of error detection; they do not describe accuracy of error detection (i.e., the proportion of errors detected). Accuracy of error detection will be addressed in Section 4.4.2.1.

The mean percentages of agreement within evaluator groups are shown in Table 13. In general, the mean percentages of agreement were slightly higher for the TEF evaluator group. They ranged from 53% to 86% for the TEF group. The checklist group obtained mean percentages of agreement ranging from 56% to 78%.

The mean percentages of agreement indicate that the use of the TEF results in slightly superior agreement with regard to error detection when compared to the traditional method of evaluation.

4.4.1.2 Variance Ratios

In order to be meaningful in the OJT environment, the use of an evaluation instrument should result in the same end score when two or more evaluators observe identical task performances. In other words, the use of an evaluation instrument should result in evaluator agreement regarding end scores derived. Large amounts of variance in end scores indicate evaluator disagreement. Smaller variances indicate evaluator consistency in end score assignments. The variance ratios compare the amount of variance in end scores obtained by the checklist group with the end score variance obtained by the TEF group. Significant variance ratios indicate that the evaluator group represented by the variance in the denominator (the TEF evaluator group) had less variance with regard to end score assignments. Non-significant variance ratios indicate that there was no difference in consistency of end score assignments between the two evaluator groups.

Less variance was expected in the TEF end scores since end scores are automatically assigned based upon the errors detected. The checklist evaluators assigned end scores based upon observed performance but without specific guidance regarding scoring task performance.

The range of end scores, the variance ratios, and significance levels are shown for each of eight task performances in Tables 14 and 15. The TEF evaluator group had significantly less variance in end scores for seven of the eight task performances. This is illustrated by the smaller range in end scores and significant variance ratios for these task performances. There was no difference in end score variance for the low performance of the Handcuff task.

The variance ratios indicate that the use of the TEF results in less variance in end score assignment when compared to the checklist method of evaluation. In fact, for three of the "bad" task performances, the use of the TEF resulted in no variance in end score assignment. All of the TEF evaluators scored an automatic failure for those task performances. The reduction in end score variance was expected with the TEF because the TEF scoring criteria guarantee that when the same errors are observed, identical scores will be derived.

Table 13. Mean Percentages of Agreement Between Evaluators for Error Detection

Preflight

	Good Performance	Bad Performance
TEF	77	84
Checklist	69	58

Refuel

	Good Performance	Bad Performance
TEF	66	86
Checklist	59	78

Handcuff

	Good Performance	Bad Performance
TEF	53	82
Checklist	56	62

M-16

	Good Performance	Bad Performance
TEF	59	81
Checklist	57	76

Table 14. Range of End Scores and Variance Ratios

Preflight Good

	Range of End Scores	Variance Ratio	Significance Level
TEF	88 - 92	29.89	p = < .01
Checklist	30 - 70		

Preflight Bad

	Range of End Scores	Variance Ratio	Significance Level
TEF	0	Undefined	n/a
Checklist	30 - 70		

Refuel Good

	Range of End Scores	Variance Ratio	Significance Level
TEF	84 - 100	12	p = < .01
Checklist	30 - 90		

Refuel Bad

	Range of End Scores	Variance Ratio	Significance Level
TEF	0	Undefined	n/a
Checklist	0 - 70		

Table 15. Range of End Scores and Variance Ratios

Handcuff Good

	Range of End Scores	Variance Ratio	Significance Level
TEF	92 - 99	52.41	p = < .01
Checklist	20 - 80		

Handcuff Bad

	Range of End Scores	Variance Ratio	Significance Level
TEF	41 - 93	1.54	Not Significant
Checklist	10 - 50		

M-16 Good

	Range of End Scores	Variance Ratio	Significance Level
TEF	92 - 99	75.70	p = < .01
Checklist	30 - 90		

M-16 Bad

	Range of End Scores	Variance Ratio	Significance Level
TEF	0	Undefined	n/a
Checklist	10 - 50		

4.4.1.3 Percentage of Pass/Fail Assignments

OJT supervisors often require GO/NO-GO decisions regarding task performance. GO/NO-GO decisions (i.e., pass/fail decisions) indicate whether or not the task was successfully performed. An evaluation instrument should result in consistent pass/fail decisions. In other words, when two or more evaluators observe identical task performances, the same pass/fail decision should be assigned.

The percentages of passes and fails each group assigned to the task performance are shown in Table 16. Examination of the table reveals that with the exception of one evaluator of the "bad" performance of the Handcuff task, the use of the TEF consistently resulted in 100% evaluator agreement for pass/fail decisions.

The percentages of pass/fail assignments indicate that use of a TEF results in superior evaluator agreement regarding whether or not the task was successfully performed.

4.4.2 Validity Results

The results of the following analyses are included:

1. T-test: Error Detection.
2. T-test: Discrimination in End Score Assignment.
3. Chi-square: Discrimination Pass/Fail Decisions.
4. Correlations: Errors and End Scores.

4.4.2.1 T-Test: Error Detection

Measures of accuracy of error detection indicate whether an evaluation instrument can be used by evaluators to detect errors in task performance. The t-tests compared the number of errors detected by the TEF group with the number of errors detected by the checklist group. Significant positive t's indicate superior error detection by the TEF evaluator group; significant negative t's indicate superior error detection by the checklist evaluator group.

The mean proportion and mean numbers of errors detected, the t's, and significance levels are shown for each task performance in Table 17. Examination of the table reveals little difference between evaluator groups in the number of errors detected for the Refuel and Preflight task performances. Although there was a tendency for the TEF group to detect more errors for the M-16 and Handcuff task performances, only one of the t's reached significance. The use of the TEF resulted in superior detection of errors for the "good" performance of the M-16 task (t = 2.86, p = < .05). None of the other t-tests approached significance.

Table 16. Percentage of Pass/Fail Assignments

	TEF		Checklist	
Performance	Pass	Fail	Pass	Fail
Preflight "good"	100	0	33	67
Preflight "bad"	0	100	33	67
Refuel "good"	100	0	17	83
Refuel "bad"	0	100	17	83
Handcuff "good"	100	0	60	40
Handcuff "bad"	10	90	20	80
M-16 "good"	100	0	40	60
M-16 "bad"	0	100	0	100

Table 17. Accuracy of Error Detection

Task Performance	Total Errors	Proportion TEF	$\bar{X}$ Proportion Checklist	$\bar{X}$ # TEF	$\bar{X}$ # CL	t	Significance Level*
Preflight Good	3	88	83	2.63	2.5	.48	Non-significant
Preflight Bad	6	69	67	4.13	4.0	.17	Non-significant
Refuel Good	2	56	67	1.13	1.38	.72	Non-significant
Refuel Bad	6	90	92	5.38	5.5	.48	Non-significant
Handcuff Good	3	67	47	2.1	1.4	1.7	Non-significant
Handcuff Bad	4	90	68	3.6	2.7	2.24	Non-significant
M-16 Good	3	67	37	2.0	1.1	2.86	$p = < .05$
M-16 Bad	6	85	68	5.1	4.2	2.24	Non-significant

*Degrees of freedom = 14.

The results of the t-tests for the difference between the number of errors detected indicate that, in general, the two evaluation methods result in equal accuracy regarding error detection. However, examination of the mean proportion and mean number of errors detected reveals a tendency for the TEF to result in superior error detection for the Security Police tasks. These results indicate that the TEF method of performance evaluation is at least as good as existing methods with regard to error detection.

4.4.2.2 T-Test: Discrimination in End Score Assignment

Discrimination in end score assignment represents the ability of the evaluators to discriminate between high and low levels of task performance with regard to the end score derived. In other words, discrimination in end score assignment reflects the ability of the evaluators to assign higher end scores to the "good" performance of each task and lower end scores to the "bad" performance of each task.

A significant positive t indicates that the mean end score derived for the "good" performance was significantly higher than the mean end score obtained for the "bad" version of the task.

Specifically, a significant difference between the means of the "good" and "bad" task performances indicates that the evaluators subtracted more points for the critical errors which occurred in the "bad" performances compared to the number of points subtracted for the minor errors inserted in the "good" performances.

The amount of points subtracted per error is predetermined in the TEF evaluation method. In the traditional evaluation method, no guidance is provided with regard to how the detected errors should affect the end score.

The mean end scores, t's, and significance levels are illustrated for each task in Table 18.

When the difference between mean end scores derived by the TEF evaluator group was tested, significant t's were obtained for all four tasks.

The difference between the mean end scores derived by the traditional evaluator group was significant for three of four tasks. The traditional evaluator group failed to discriminate between performance levels of the Preflight task. In fact, the traditional evaluator group assigned slightly higher end scores to the "bad" version of the Preflight task ("bad": $\overline{X} = 54$, "good": $\overline{X} = 48$).

Even though the traditional evaluation method resulted in significant differences in end scores for three of the tasks, the use of the TEF consistently resulted in a larger difference between the scores for the "good" and "bad" task performances.

Table 18. T-Tests for the Difference Between the Mean End Scores of Good and Bad Performances

Preflight	$\bar{X}$ End Score Good Performance	$\bar{X}$ End Score Bad Performance	t*	P
TEF	89	0	118.38	p = < .01
Checklist	48	54	-.89	non-significant

Refuel	$\bar{X}$ End Score Good Performance	$\bar{X}$ End Score Bad Performance	t*	P
TEF	89	0	42.83	p = < .01
Checklist	61	25	3.18	p = < .05

Handcuff	$\bar{X}$ End Score Good Performance	$\bar{X}$ End Score Bad Performance	t†	P
TEF	94	46	9.03	p = < .01
Checklist	59	32	4.99	p = < .01

M-16	$\bar{X}$ End Score Good Performance	$\bar{X}$ End Score Bad Performance	t†	P
TEF	95	0	125.25	p - < .01
Checklist	59	19	4.99	p = < .01

*Degrees of freedom = 14

†Degrees of freedom = 18

Comparison of the range of mean end scores obtained by the two evaluator groups reveals that the TEF group consistently assigned relative higher scores to the "good" performances and relatively lower scores to the "bad" performances.

For the "good" task performances, the TEF evaluator group obtained mean end scores ranging from 89 to 95 compared to mean end scores ranging from 0 to 46 for the "bad" performances. The difference in mean end scores obtained by the traditional evaluators was not as great; traditional evaluators obtained mean end scores ranging from 48 to 61 for the "good" task performances and 19 to 54 for the "bad" task performances.

Compared to the TEF group, the checklist evaluators subtracted more points for the minor errors inserted in the "good" task performances and less points for the serious errors included in the "bad" versions of the task performance.

In summary, the results of the t-tests indicate that use of the TEF consistently results in end scores which discriminate between "good" and "bad" levels of performance. On the other hand, use of the checklists resulted in discrimination for three of the four tasks evaluated.

Examination of the range of end scores derived revealed a consistent tendency for the use of TEFs to result in better performance-level discrimination. In addition, the end scores reflected a tendency of the TEF evaluator group to derive scores which reflected the amount and type of errors detected.

4.4.2.3 Chi-Square: Discrimination of Pass/Fail Decisions

The previous discussion focused upon discrimination between performance levels with regard to end score. In order to be useful in the OJT environment, a performance evaluation instrument should also result in discrimination between performance levels with regard to pass/fail decisions. Use of the evaluation instrument should result in the assignment of pass decisions to the "good" task performance and fail decisions to the "bad" task performance.

A significant positive chi-square indicates that discrimination between performance levels occurred. Non-significant chi-squares indicate that there was no difference in the number of pass/fail decisions assigned to the "good" and "bad" performances.

The contingency charts in Table 19 illustrate, by evaluator group, the pass/fail decisions assigned to the "good" and "bad" performances of each task. The chi-squares and significance levels are shown under each contingency chart.

Table 19. Chi-Squares Pass/Fail Decision by Performance Level

		TEF		Checklist	
		Good	Bad	Good	Bad
Preflight	Pass	8	0	2	2
	Fail	0	8	6	6
		$X^2 = 12.25$		$X^2 = 0$	
		$P = < .01$		P = Non-significant	
Refuel	Pass	8	0	1	1
	Fail	0	8	7	7
		$X^2 = 12.25$		$X^2 = 0$	
		$P = < .01$		P = Non-significant	
Handcuff	Pass	10	1	6	2
	Fail	0	9	4	8
		$X^2 = 16$		$X^2 = 3.3$	
		$P = < .01$		P = Non-significant	
M-16	Pass	10	0	4	0
	Fail	0	10	6	10
		$X^2 = 200$		$X^2 = 5$	
		$P = < .01$		$P = < .05$	

Observation of Table 19 reveals that use of the TEF resulted in almost perfect discrimination, with the exception of one evaluator who assigned a pass to the "bad" performance of the Handcuff task. The TEF evaluators always assigned a pass to the high performance and assigned a fail to the low performance. The ability of the TEF evaluators to discriminate between performance levels is reflected in the significant chi-squares obtained for all four tasks.

The use of the traditional evaluation method did not result in performance-level discrimination in pass/fail decisions. Only the chi-square for the M-16 task was significant. In fact, for the two Aircraft Maintenance tasks, the traditional evaluators assigned equal numbers of pass and fail decisions to the high and low performance levels, resulting in chi-squares of 0.

In summary, the pass/fail decisions assigned by TEF evaluators accurately reflected observed performance. The TEF can be used successfully in the OJT environment to make pass/fail decisions. However, the effectiveness of the use of traditional evaluation methods to make pass/fail decisions is questionable. The pass/fail decisions resulting from the traditional evaluation method did not discriminate with regard to performance level for three of the four tasks. On the other hand, the TEF evaluators consistently discriminated between performance levels, indicating "pass" for the "good" performances and "fail" for the "bad" performances.

4.4.2.4 Correlation: Errors and End Scores

The relationship between the end scores and overall level of performance was discussed in Section 4.4.2.1. In addition to discriminating between "good" and "bad" performance levels, end scores should also reflect the amount and type of errors which occur in task performance. The end scores should vary according to the amounts and types of errors detected.

The correlations between the end scores assigned and the errors detected indicate the sensitivity of the scores to subtle changes in performance. Positive correlations indicate that as the number of detected errors increased, the end scores derived decreased. Negative correlations indicate that as the number of detected errors increased, the end scores decreased. Negative correlations result when the number of points subtracted from 100 (the score for perfect performance) increases as the number of detected errors increases. Perfect negative correlations were not expected since errors of varying degrees of criticality were inserted in the task performances.

The correlations and significance levels are shown in Table 20. Evaluations with the TEF resulted in significant negative correlations for five of the task performances. (Only five correlations were calculated for the TEF results, due to restriction of range in the end

Table 20. Correlation of End Scores With Number of Errors

	TEF		Checklist	
	r	p	r	p
Preflight Good	-1.00	p = < .01	+.23	Non-Significant
Preflight Bad	*		-.65	p = < .05
Refuel Good	-.83	p = < .01	-.41	Non-Significant
Refuel Bad	*		+.32	Non-Significant
Handcuff Good	-.78	p = < .01	+.17	Non-Significant
Handcuff Bad	-.69	p = < .05	-.48	Non-Significant
M-16 Good	-.84	p = < .01	-.50	Non-Significant
M-16 Bad	*		-.76	p = < .01

*Correlations were not computed for these performances since all of the evaluators assigned end scores of 0.

scores of the remaining three performances.) The correlations ranged from -.69 for the "bad" performance of the Handcuff task to a perfect negative correlation (-1.0) for the "good" performance of the Preflight task. Only two of the eight task performances evaluated by the checklist evaluators obtained significant negative correlations. The correlations for "bad" performances of the Preflight task ($r = -.65$, $p < .05$) and the M-16 task ($r = -.76$, $p = < .01$) were significant.

The correlations indicate that the use of the TEF consistently results in end scores which vary according to the number of errors detected in task performance (i.e., when more errors are detected, more points are subtracted resulting in a lower end score). On the other hand, end scores resulting from the traditional method of evaluation were not always related to the number of errors detected in task performance. The end scores did not decrease as the number of errors increased.

The end scores which result from a TEF evaluation are, by definition, sensitive to the number and type of errors which occur in task performance since points are preassigned to all potential errors. On the other hand, the checklist method provides no guidance to the evaluator regarding the amount of points to be subtracted per error.

4.5 Summary and Conclusions

Evaluations of OJT performance using the Task Evaluation Forms were substantially more reliable and valid than evaluations using the traditional checklist method. The TEF method was superior to the checklist method for almost every measure of reliability and validity assessed in this study.

The results of the reliability analysis of the evaluation methods indicate:

1. Consistency of error detection between evaluators is slightly superior for the Task Evaluation Form method.

2. Consistency of end score assignment between evaluators is clearly superior when the TEF is used. The nature of the scoring criteria of the TEF guarantee that when the same errors are detected by more than one evaluator, the same score will be derived. Evaluator subjectivity in assigning a score to a task performance is eliminated when the Task Evaluation Form is used.

3. Evaluator disagreement with regard to pass/fail decisions is also eliminated with the TEF. The TEF evaluators demonstrated a greater degree of consistency in pass/fail assignments when compared to the checklist evaluators.

The validity analysis of the evaluation methods indicates:

1. There is no significant difference between the two methods in error detection. The TEF is at least as useful as the checklist in its ability to guide evaluators in detecting errors.

2. The TEF is superior in discriminating between high and low levels of performance. The end scores resulting from the TEF evaluations discriminated between the high and low levels of performance for all tasks. However, the checklist evaluator group assigned slightly lower scores to the higher level of performance of the Preflight task. The ability to discriminate between levels of performance is an important characteristic of any evaluation method. It allows comparison between performers and comparison of different performances by the same performer.

3. The TEF is superior to the checklist in providing criteria for assigning a pass/fail decision to task performance. The TEF group consistently assigned a "pass" to the higher-level performance and a "fail" to the lower-level performance. On the other hand, the checklist evaluators assigned an equal number of passes and failures to both performances of three of the four tasks. The traditional method provides no criteria for making a pass/fail decision. The decision is left entirely to the individual evaluator's discretion.

4. TEF evaluations result in end score decisions which accurately reflect performance level. The end scores vary according to the amount and type of errors detected. This relationship of end score with errors detected does not occur consistently when the checklist method is used.

The results indicate that the TEFs can be used to provide guidance to the evaluator in the following ways:

1. Detecting errors in task performance.

2. Assigning a reliable end score based on the amount and type of errors detected.

3. Making a reliable pass/fail decision based on the amount and types of errors detected.

4. Reliably comparing or discriminating between levels of performance.

The major strength of the TEF is its ability to provide guidance to the evaluator in interpreting the results of an evaluation. The checklist method of evaluation does not provide criteria for scoring an evaluation, making a pass/fail decision, or comparing performers based on their evaluations. When the TEF is used to evaluate performance of OJT tasks, evaluator subjectivity in the assignment of end scores and pass/fail decisions is eliminated. The results of a TEF evaluation actually reflect the amount and type of errors detected, thus allowing comparison between observed levels of performance.

5.0 GENERALIZABILITY ASSESSMENT

Previous analyses demonstrated the usefulness of the TEF development procedures in two very different career fields: Aircraft Maintenance and Security Police/Law Enforcement. These two career fields included tasks which were equipment oriented and non-equipment oriented. However, it was felt that these two areas were not representative of all Air Force jobs. Additional study was needed to demonstrate the generalizability of the procedures.

For this purpose, the Personnel AFSC was chosen for assessment. The Personnel area was expected to contain tasks very different from those in the other areas considered. The three career fields, taken together, provide a better sample of job and task types within the Air Force environment.

The Personnel assessment was designed to demonstrate the applicability of the TEF development procedures to tasks in this career field. The sections below describe the applicability criteria, the research approach, the data collection and analysis procedures, and the findings of the generalizability assessment.

5.1 Objectives

To determine if the TEF development procedures are applicable to the Personnel field, it was necessary to determine if the Personnel AFSC and its tasks have certain characteristics. The following criteria had to be met for the procedures to be applicable:

1. It must be possible to divide the duties and responsibilities into specific tasks.
2. It must be possible to describe beforehand the steps that must be performed to complete a task.
3. It must be possible to describe the performer's actions and task outcomes that indicate successful task performance.
4. The TEF Evaluation Areas must be relevant to the critical performer actions and task outcomes.
5. At least one of the possible methods of TEF utilization must be feasible.

The first two of these criteria relate to the nature of the tasks within the AFSC and whether they can be described for evaluation purposes. The third and fourth criteria relate to the feasibility of evaluating these tasks with the combined critical incident technique and logical analysis approach method in general, and with TEFs in particular. The final criterion relates to whether TEFs can be used in the Personnel OJT environment.

5.2 Overview of Approach

Several activities were accomplished to determine the applicability of the TEF development procedures to tasks in the Personnel area:

1. Review of Personnel regulations.
2. SME interviews.
3. Preliminary application of procedures.
4. Handbook revisions.
5. Trial TEF development.

Each of these activities is described below.

5.3 Data Collection and Analysis

Personnel regulations were reviewed to determine the general nature of the duties and responsibilities within the Personnel career field. This review focused on the criteria for TEF applicability, especially in terms of task types. That is, emphasis was placed first on determining the proportion of duties and responsibilities that could be divided into well-defined tasks. Secondly, for the identified tasks, consideration was given to the proportion of tasks which could be described beforehand in a step-by-step manner.

SMEs from the Consolidated Base Personnel Office (CBPO) at Bergstrom Air Force Base, Texas, were selected for interview. The SMEs represented seven different CBPO work centers and their combined experience covered 14 work centers.

The SMEs were interviewed separately regarding the duties and responsibilities in their respective work centers. Following the interviews, the SMEs reviewed the TEF development procedures individually and as a group.

The following issues were addressed during the SME interviews and TEF review:

1. What proportion of work center duties and responsibilities can be divided into tasks?
2. Can the steps in the task be described beforehand?
3. What types of performer actions and task outcomes should be evaluated in the Personnel AFSC?

4. How should "criticality" be defined for the Personnel AFSC?

5. What types of OJT evaluations are currently conducted?

As part of the preliminary data collection, the SMEs were asked to apply the TEF development procedures to a set of Personnel tasks. This preliminary application of the procedures focused on the following tasks:

1. Preparing Request for Manning Assistance.

2. Mobility Passport Processing.

3. Writing a Simple Direct English Statement Information Retrieval System (DESIRE).

4. Preparing Request for Designated Move of Dependents to Foreign Country.

Based on the review of regulations, SME interviews, and preliminary application of the procedures, an interim handbook was developed. This handbook included examples from Personnel tasks and was designed to apply specifically to the types of tasks found in the Personnel career field.

After the interim Personnel handbook was developed, the same group of SMEs were again asked to apply the TEF procedures. For this trial TEF development, each SME selected two tasks performed in their work center: one which was straightforward and another which required the development of a task scenario. TEFs were developed for the following tasks:

1. Daily Startup of Automated Personnel Data System (APDS) II.

2. Removing a Projected Promotion (Based on Non-Recommendation, Lower Grade Airman).

3. Final Outprocessing Continental United States to Continental United States, no Assignment Instruction Codes (CONUS to CONUS with No AICs).

4. Completion of Individual Mobilization Augmentee (IMA) Folders.

5. Retraining Application.

The findings of this data collection effort are presented below.

5.4 Results

The findings of the generalizability assessment were positive, showing that the TEF development procedures can be applied to a broad range of tasks with good results. As noted previously, several criteria had to be met to demonstrate applicability of the procedures to Personnel tasks. These criteria related to task types, defining successful performance, and the evaluation environment.

The first task-related criterion was that it must be possible to divide the AFSC's duties and responsibilities into tasks. For the Personnel field, this criterion was met by the majority of duties and responsibilities. The second task-related criterion was that it must be possible to describe the steps in a task before performance. For many tasks in the Personnel AFSC, task description required the creation of a task scenario. In other words, a simple task title (e.g., "Final Outprocessing") could not be translated into task steps without extensive use of conditional statements. To describe a straightforward list of steps, it was necessary to add more specifics to the task title (e.g., "Final Outprocessing, CONUS to CONUS, no AICs"). The development of such task scenarios is provided for in the TEF development procedures.

The criteria related to definition of successful performance were also met. Critical performer actions and task outcomes could be identified, and these could be described in terms of the TEF Evaluation Areas. Some additions and clarifications in the handbooks were indicated. With these handbook revisions, the procedures should be easily applicable to Personnel tasks.

The revisions are listed below:

1. An explanation of how to define a task involving interactions with other offices or agencies was added.
2. Additional instructions on defining task scenarios were added.
3. The criticality questions were revised.
4. Instructions on describing paper end products such as folders or forms were added.
5. The Tools, Equipment, and Materials Use area was revised to include the use of "resources."
6. A list of variables and factors that might affect the way tasks are performed was added.

The final criterion related to utilization of the TEFs. The SMEs interviewed indicated that they do not currently have a formal OJT program or evaluations of OJT task performance. However, no obstacles to the use of the TEF in the Personnel work environment were found. The SMEs did express concern about the amount of time necessary to conduct "over-the-shoulder" evaluations, since some tasks in the Personnel career field require days or weeks to complete. The TEFs are intended to be used in an "over-the-shoulder" evaluation situation. An exception to the "over-the-shoulder" method is indicated when tasks require days or weeks to complete and only end products are evaluated. Thus, those Personnel tasks which require days or weeks to complete and require only the evaluation of end products can be evaluated when task performance has been completed.

5.5 Conclusion

In summary, the TEF development procedures were found to be applicable to tasks within the Personnel AFSC. This assessment, combined with previous analyses, shows that the procedures can be applied to tasks of many types from widely varying AFSCs. The TEF development procedures have shown a high level of generalizability and should be applicable to most job and task types within the Air Force OJT environment.

6.0 CONCLUSIONS

The TEF development procedures were designed to meet specific procedural and operational requirements. These requirements were presented in Sections 2.3.1 and 2.3.2, respectively. The development procedures and the TEFs themselves had to be practical and usable in the Air Force OJT environment. The TEF system, it is felt, meets all requirements. Studies performed to demonstrate its reliability, validity, and generalizability, in both procedural and operational terms, were universally positive in their results. Minor problems were resolved through an extensive, multi-level cycle of revisions. The features of the TEF development procedures, compared to the initial requirements, are presented in detail below.

6.1 Procedural Validity

Requirement: The development procedures must result in the identification of those aspects of task performance which are directly related to successful task performance.

The results of the procedural validity assessment described in Section 3.0 demonstrated the validity of the TEF development procedures. The TEFs resulting from procedural application were consistently high with regard to the accuracy and criticality of the information included. Initially, the completeness of the forms was questionable; however, the completeness was improved through revisions in the presentation of the development procedures. Thus, TEFs depict, for the evaluator, behaviors and outcomes which are critical to successful task performance.

6.2 Procedural Reliability

Requirement: The development procedures must be consistently applied by two or more users.

Once again, the results of the procedural assessment demonstrated adequate reliability in both career fields. The TEF development procedures can be applied by SMEs to identify:

1. Critical evaluation areas related to task performance.

2. Critical events within an evaluation area.

3. Standards or criteria for evaluating these events.

4. When in the task these events should be evaluated.

One of the goals of the procedural assessment was to target specific problem areas in the presentation of the TEF development procedures. It is expected that the procedural reliability of TEFs generated in the future will be improved by enhancements which are indicated by the assessment results.

6.3 Procedural Utility

Requirement: The development procedures must have the capability of application by Air Force users.

One of the initial concepts that shaped the TEF development procedures was SME responsibility for form development. SME developers have the advantage of knowing the task, and do not require the lengthy data collection process which is necessary when a non-SME is responsible for form development. In order to keep the TEF development procedures usable by SMEs, no special knowledge of educational principles or assessment technology was assumed. Instead, detailed instructions and a common-sense strategy allow the SME to apply task performance experience to form development. At each step, items of information from the SME's own knowledge are recorded and sorted. As an end result, key elements of task performance have been identified. Several assessments of the TEF system included trial form development by SMEs. The final handbooks were influenced by SME responses and criticisms at these times. The TEF development procedures, it is felt, can be successfully used by SMEs with a minimum of training and orientation.

6.4 Procedural Generalizability

Requirement: The assessment instrument development procedures must be applicable to tasks from all specialty areas, including both maintenance and non-maintenance specialty areas.

The TEF development procedures have been shown to be widely generalizable across both tasks and career fields. On the task level, TEFs can be produced for many different types of tasks. Trial TEF development efforts successfully produced valid forms for maintenance and non-maintenance tasks, for equipment-oriented and non-equipment-oriented tasks, for complex operational checks and straightforward remove-and-replace tasks, and for tasks involving paperwork. In addition, the TEFs can be generated for frequently performed tasks, as well as tasks requiring a rigged task scenario for evaluation purposes.

TEFs can be used to evaluate task performance for many AFSCs as well as a broad range of task types within AFSCs.

6.5 Operational Validity

Requirement: The results must accurately reflect the number and types of errors which occur in task performance.

In order for the results of TEF evaluations to be meaningful, it must be demonstrated that the TEFs do, in fact, serve their intended purpose (i.e., assess task proficiency). In other words, a certain relationship must exist between the evaluation results and the observed performance. In the TEF methodology, specific numbers of points are assigned to each potential error in task performance. The score assigned to a task performance accurately reflects the number and types of errors which occurred in that task performance. The assignment of a task fail decision is based upon the end score obtained. Thus, pass/fail assignments and end scores resulting from a TEF evaluation have the advantage of being clearly related to performance, which makes them optimally meaningful in the operational context.

6.6 Operational Reliability

Requirement: Two or more evaluators observing the same task performance should derive identical results.

Operational reliability is important in order for the evaluation results to be useful in the OJT environment. The results of the operational assessment discussed in Section 4.0 clearly demonstrate the reliability of the pass/fail assignments and end scores resulting from TEF evaluations. This high degree of reliability allows comparisons between results obtained by different evaluators. The reliability (i.e., standardization) of the evaluation scoring criteria is related to the overall specification of the evaluation situation. The implications of increased standardization are discussed in the following paragraphs.

6.7 Operational Standardization

Requirement: Two or more evaluators conducting an evaluation of the same task should conduct the evaluation in the same manner.

Use of the TEFs will standardize task assessment on several levels. First a given evaluator will assess each performer in the same way. Without a standardized form, OJT instructors (or other evaluators) may unintentionally consider different factors when evaluating different performers. A TEF directs the evaluator's attention to the same aspects of performance for each assessment.

Standardization will also occur on the OJT program level. That is, each evaluator concerned with a particular task will evaluate that task with the same form. Without such a standard methodology, different evaluators may consider widely varying factors. Using TEFs for task evaluations throughout a program will ensure that one performance standard is applied.

Finally, using the TEFs will promote standardization across programs. The same type of standard will be applied to all tasks evaluated with TEFs. Thus, the scores found in one OJT program will be on the same scale as those from other programs. Standardization on this level allows for comparisons among OJT programs as well as between students.

6.8 Operational Utility

Requirement: The results of evaluations must provide direct feedback regarding specific deficiencies in task performance. In addition, varying levels of detail must be provided so that the user can select the appropriate level for the intended purpose.

The final requirement is that the forms and results derived from evaluations with those forms are operationally useful.

The TEFs are designed to be useful in the operational environment. The forms are simple and easy for the evaluators to use. Evaluators' responsibilities are minimized. The evaluator simply records performance errors and subtracts the points related to each error to obtain the end score.

The results of the TEF evaluation indicate that the TEFs are operationally useful. The standardization of the evaluation situation and the high reliability and validity of the TEF evaluation results optimize the usefulness of those results in the OJT environment. The end scores and pass/fail assignments provide information for certification as well as allow direct comparison between trainees, units, or OJT programs. The numerical end scores provided by the TEF are on a standard 0 to 100 scale. This optimizes their clarity and meaningfulness, as this type of scale is familiar to most individuals.

In addition, the TEFs provide more than a pass/fail decision and numerical score. The TEF evaluation results actually describe observed performance. A separate score is derived for each evaluation area, and specific errors in task performance are identified. This detailed information allows the evaluator and the trainee to see what types of errors were made (in terms of evaluation areas or steps). This detailed information may not be needed, in which case only the pass/fail assignment or overall end score would be used. In some cases, however, it is expected that the detailed information will be useful. One use for detailed scoring is performance interpretation. Not only can performance be categorized as successful or unsuccessful, but the areas of correct and incorrect performance can also be pinpointed. This capability has obvious potential for improving feedback to students. The areas where students have difficulties can be identified for special attention. Additionally, such information could benefit training design efforts, as specific deficiencies in an OJT program could also be identified.

6.9 Summary

The TEF methodology clearly meets all of the requirements set forth at the beginning of this effort. The TEFs have potential for several uses within the OJT environment including: task certification, performance tracking, identification of individual or unit deficiencies, determination of unit readiness, and performance comparisons on an individual or unit level. Consideration should be given to the utilization of the TEF development procedures and the resulting instruments. Recommendations for the integration of the TEF methodology into the OJT environment are presented in the next section.

7.0 RECOMMENDATIONS

Study of on-the-job performance evaluation methodology and approaches in this series of efforts has concentrated upon development of a valid, reliable, and easy-to-use methodology for which no special expertise in measurement theory or assessment techniques is necessary. This objective having been accomplished, application of the TEF methodology in practice appears to be a logical next step in attaining improved assessment capability for Air Force OJT. One potential milieu for application of this methodology may be the Advanced On-the-Job Training System (AOTS) currently under development by AFHRL. A number of considerations must be taken into account in deciding upon the most appropriate applications of the TEF methodology. The following recommendations are provided to support those considerations.

7.1 AFSCs for TEF Application

Selection of appropriate AFSCs for development and implementation of the TEF evaluation process will be extremely important to the acceptance of the methodology and its ultimate success or failure. There are two classes of AFSCs for which TEF adoption may be considered. These are:

1. AFSCs which presently possess OJT trainee performance evaluation systems.
2. AFSCs which currently possess no OJT evaluation or measurement capability, but where such capability is needed.

In the first case, the current evaluation systems for OJT trainee performance may not provide the results that are sought in terms of skill and ability diagnosis, utility, validity, or reliability of the methodologies. TEF adoption should certainly be considered an option if these conditions continue. In the second case, careful judgment about adopting the TEF measurement methodology must be exercised, so that the evaluation system is implemented appropriately and effectively. Some considerations in selecting AFSCs for adoption of the TEF OJT performance evaluation approach are:

1. AFSCs selected for TEF methodology adoption should, in general, be ones wherein performance of job tasks results in tangible, observable outcomes, rather than intangible, judgmental, or conceptual outcomes. An AFSC where job performance and performance "products" are not observable or measurable in some way is inherently inappropriate for an operationally oriented measurement methodology such as the TEF approach.

2. AFSCs where tangible products are the main criterion of job performance success will be less appropriate candidates for TEF adoption than those where the process of task performance is of major importance, in addition to the production of tangible products. Although the TEF methodology can be applied to specialties where products are the principal criterion of good job performance, such an application would be inappropriate. The TEF approach centers on the process of performance as a measurement objective with products of performance as incorporated criteria for good performance. In general, in specialties where a particular product or set of product characteristics is much more important than the process of attaining the product, then TEFs should not be developed or used. Doing so would be a needless and costly expenditure of scarce resources.

3. In general, the TEF approach is an appropriate approach to adopt for those specialties where a majority of the evaluation dimensions incorporated in the TEF methodology are appropriate to judging the successful or unsuccessful performance of tasks. It is recommended that any specialty under consideration be looked at from a global point of view, in terms of the general types of tasks performed by personnel in the specialty, in order to make this judgment. If a majority of tasks appear to relate to only two or three of the evaluation dimensions, then a limited TEF development (excluding the inapplicable or inappropriate dimensions) might be considered. In cases where performance criteria are restricted to only one of the TEF evaluation dimensions, TEF development should not be considered.

7.2 Selecting Tasks Within Specialties for TEF Development

Just as not all specialties will be appropriate for use of the TEF evaluation approach, not all tasks within a given specialty will be suitable for development of TEFs. Although the TEF development procedures contained in the handbooks do not provide procedures for selecting tasks for TEF development (this is assumed to be a process external to TEF development), the following criteria are recommended for consideration in task selection for TEF development.

7.2.1 Frequency of Task Performance

In general, tasks that are performed with high frequency are more appropriate for TEF development than those which are performed infrequently. This is the case for two reasons. First, evaluation opportunities for on-the-job, over-the-shoulder evaluation will be more frequent with frequently performed tasks. This means equipment will not have to be rigged, or simulation of the task utilized, to conduct performance assessments. This lowers the overall effort of evaluation -- an important factor in the acceptance of an evaluation method in busy, readiness-oriented units. Second, tasks that are higher in frequency will have more training effort associated with them than those of lower frequency, in general. It is probable that more individuals will be required to perform the frequent tasks, thus requiring more people to be trained and evaluated on task performance.

7.2.2 Task Criticality

Tasks whose successful performance is necessary to maintain readiness or mission capability should be favored for TEF development over less critical tasks. Although practically all tasks are critical in some sense, those which are most closely related to mission capability or readiness are those which are probably the most important from the OJT and evaluation point of view, since they are the tasks that must be completed effectively. Task criticality information is commonly available if Instructional Systems Development (ISD) procedures have been used to develop training for a particular specialty, since criticality is one of the judgment factors used in selecting tasks to be trained.

7.2.3 Proportion of Personnel in Specialty Who Perform the Tasks

In general, tasks which are performed by large proportions of personnel in a given specialty should be favored for TEF development over those which are performed by smaller proportions of personnel, other factors being equal. The larger the proportion of personnel who perform a task, the greater is the potential need for OJT and associated performance evaluation of the task. Proportion of performance information can be derived from the results of occupational surveys, which are periodically administered to many specialties by the Air Force Occupational Measurement Center (AFOMC), Air Training Command.

7.2.4 Data Availability

It is obviously infeasible to develop TEFs for tasks where performance is not reasonably well standardized, and for which data on task performance are either not available or faulty. Thus, TEFs should be developed only for tasks on which comprehensive and complete information specifying the characteristics of task performance and products (if applicable) is available.

7.2.5 Task Complexity

Complex tasks, in general, have more performance elements incorporated in them; thus, task performance assessment is more important, since many intermediate goals must be attained during task performance. Also, assessment of complex tasks is itself often complex, and the structure and guidance for evaluation provided by the TEF approach may provide more accurate and comprehensive evaluation for such tasks than less structured and less comprehensive evaluation approaches. Tasks which are complex, or have large numbers of intermediate objectives, should therefore be selected for TEF development preferentially over simple or straightforward tasks.

7.2.6 Requirements for Demonstration of Task Proficiency

This criterion relates to the need for proficiency certification for some tasks in some specialties. Frequently, tasks are considered sufficiently critical or hazard-prone in performance that certification of performance capability is required for task performers before they are allowed to perform or supervise the task. This is a critical element of the OJT system, especially in maintenance and weapons-related tasks, since certification takes place as part of the OJT program, or in parallel with OJT. Tasks for which certification is required should be considered among the strongest candidates for TEF development, since the TEF structure provides natural criteria for assessing competency at the high levels of performance required for certification.

7.3 Data Sources for TEF Development

Obviously, accur te, reliable information which describes all aspects of performance of a task is required to develop a TEF for that task. In many cases, regulations, directives, Technical Orders (T.O.s), and other documentation contain detailed descriptions of the processes required to perform a task and the products to be developed as a result of task performance. When selecting data sources to support TEF development, the following factors should be considered:

1. Recency - Generally, the most recent sources available will contain the most current and accurate information regarding task performance. Unless there is specific reason to doubt the reliability or completeness of the information in a particular source, the most recent or recently updated sources should be used to support TEF development.

2. Comprehensiveness - Complete documentation of all aspects of task performance is required for development of reliable and valid TEFs. Therefore, as wide a variety of sources as is available should be obtained and consulted by the TEF developer, to ensure that all aspects of performance are addressed in the resulting TEF. It is not unusual for information needed to completely evaluate all aspects of task performance to be distributed across a number of

sources. In preparing to develop a TEF, a thorough search for all descriptive information is critical to ensure the ultimate utility and validity of the TEF.

Existing data sources such as regulations, T.O.s, etc. should be considered as primary sources for task descriptive data. In the future, automated task analysis systems such as AFHRL's Automated Task Analysis Authoring Aid (ATA-3) may provide additional sources for task-descriptive information for use in TEF development. The TEF development automation support effort performed in the present effort is expected 'n ultimately interface directly with the data products of ATA-3 and other similar systems currently in development (some in the context of AOTS). This wi'l facilitate TEF development and updating significantly.

7.4 TEF Developer Characteristics

The most critical element in the TEF development process is the individual developer, who makes the decisions which determine TEF content. Consequently, the personnel chosen to develop TEFs must be carefully selected to ensure that their work results in a valid, reliable, usable product. Some desirable (if not essential) characteristics which may be used to select TEF developers follow.

7.4.1 Communication Skills

The goal of the TEF is to effectively communicate evaluation criteria and other essential information to the evaluator. Therefore, one requisite qualification of a TEF developer is good oral and written communication skills. Some individuals are fully qualified job performers with an enormous fund of job-related skills, but cannot effectively communicate their knowledge to others. Such individuals should not prepare TEFs. Ideally, the TEF developer will be highly literate and an effective writer who also possesses the other characteristics listed below.

7.4.2 Job Qualifications

TEF developers will have a difficult time making the requisite decisions about what and what not to evaluate unless they are fully familiar with the content of the jobs and tasks for which they develop the TEFs. Therefore, TEF developers must be fully job qualified in the specialty for which they conduct development. Possession of a 5-level qualification in the specific AFSC for which TEFs are to be developed should be considered a minimum requirement, with 7-level qualification desirable, especially for highly critical or safety-implicated tasks which will be evaluated by hands-on performance. An incidental benefit of selecting 7-level personnel as TEF developers will be their thorough familiarity with information sources about tasks in their specialty. This may lessen the burden of researching data sources to some extent, making the TEF development process somewhat less time-consuming.

7.4.3 Motivation

Developing TEFs is neither an easy nor an intrinsically rewarding task. The individual developer may be easily distracted from his/her task of TEF development due to the laboriousness of the development process. This can result in unreliable or invalid TEFs which will not support effective evaluation. Individuals should be selected as developers who understand the importance of OJT and the evaluation component of OJT in the abstract, and who are stakeholders in the successful development of the TEFs for their specialty. Supervisory personnel who have had experience as OJT trainers or supervisors tend to possess these characteristics, and should be considered as prime candidates for selection as TEF developers.

7.5 Training

As pointed out earlier in this report, training will be essential for all personnel involved in the development and use of TEFs. Three categories of personnel will be involved in the generation and application of TEF instruments Each of the three categories will require different training content and emphasis. Suggested training approaches for each of the categories are presented below.

7.5.1 TEF Developer Training

If TEF developers are selected according to the recommended criteria presented earlier, relatively little training will be required to enable the developers to produce valid, reliable, and usable TEFs. Approximately 2 days of training was found to be adequate for developers during this series of studies; this amount of training should be adequate for developers of operational TEFs, as well. One day, plus some additional study time, should be devoted to study of the appropriate TEF Developer's Handbook, to develop an understanding of the general TEF approach and to become familiar with the steps of the generation process and the worksheets, documentation, and TEF preparation requirements. The second day of training should involve generation of one or two "practice" TEFs, appropriate to the specialty for which TEFs are being developed, and critical feedback on the trainee developer's performance in developing the "practice" TEFs. This training will be relatively simple to implement, since it will require only preparation of "criterion" TEFs for each specialty, and development of a brief introductory presentation to explain the purpose and objectives of TEF developer training.

7.5.2 TEF Evaluator Training

The development of explicit TEF evaluator training was beyond the scope of the present effort, but some general guidelines for evaluator training can be provided from this experience in preparing evaluators to conduct evaluations as part of the operational validation of TEFs. The TEF is a very succinct document; i.e., it contains a great deal of information in a compact format. This means that untrained evaluators may have some difficulty interpreting the information and applying the criteria without some familiarization. It is recommended that a TEF Evaluator's Handbook be developed which explains the purpose, use, and information contained in the TEF and provides guidance on conducting TEF-based evaluations. Such a handbook would be used for self-study during qualification as an OJT examiner or evaluator. A criterion performance test should also be developed which can be used to diagnose problems in application of the TEF approach during evaluations and provide feedback to improve evaluator performance. The criterion test could also be used in qualification for OJT evaluator status, as a diagnostic/remedial tool or criterion performance examination.

7.5.3 OJT Supervisor Training

Once the TEF evaluation approach has been implemented for a particular specialty, it may become desirable for OJT supervisors to use the TEFs as information resources, in addition to the usual technical job documentation (e.g., T.O.s, regulations) available. The information in a TEF provides an explicit description of acceptable performance on the job for a particular task in a very concise manner; this makes use of the TEFs to support training an attractive prospect for the OJT supervisor. Some training in the correct interpretation of TEF information may be desirable for the supervisor if the TEF is utilized in this manner. Such training might consist of review of the Evaluator's Handbook to be developed, to ensure that the information is interpreted correctly. This training could be completely informal, consisting of self-study of the handbook prior to attempts to use the TEFs as information resources. If this use of TEFs is permitted, explicit guidance that the TEF cannot substitute for official documentation, but must only be used as a summary or supplement in addition to (e.g.) T.O.s, should be provided.

7.6 TEF Integration into the OJT System

The TEF approach to OJT performance evaluation has been demonstrated in these studies to be a valid, reliable, usable, and acceptable means of assessing trainee task performance. At present, however, the TEF approach is not a part of the official Air Force OJT system. While TEFs can provide comprehensive information on trainee performance, their introduction and integration into OJT practice must be performed with care and insight. In some cases, the TEF approach may provide more information than is desired or needed by OJT supervisors or evaluators to make training decisions. Considerably more than a GO/NO-GO judgment is provided when TEFs are used as designed. This detailed level of information is quite suitable for tracking the performance of trainees and identifying training needs and deficiencies. However, the present OJT system is not constituted to make use of this detailed information.

A promising arena for introduction of the TEF evaluation approach may be AOTS, which is currently under development under AFHRL sponsorship. AOTS is intended to provide an advanced, comprehensive milieu for the conduct and management of OJT. One of the designed capabilities of AOTS is the ability to closely track and monitor OJT trainee progress and performance, and provide flexible, tailored training suited to trainee mastery and level of accomplishment. The TEF evaluation approach is capable of providing the detailed and comprehensive data needed to support this flexible and adaptive approach to OJT. The TEF approach will provide the performance data needed to diagnose and remediate trainee performance, while the AOTS management system will provide recordkeeping and scheduling for the administration of evaluations, as well as developing detailed and comprehensive training prescriptions for trainees. It is recommended that the TEF approach be seriously considered for implementation as the evaluation and performance assessment component of AOTS, when integration of such a capability is appropriate.

REFERENCES

Warm, R., & Roth, J. T. (1986, May). Task evaluation form: Development procedures for maintenance and equipment-oriented tasks (AFHRL-TP-85-55). Lowry AFB, CO: Training Systems Division, Air Force Human Resources Laboratory. (AD-A167 597)

Warm, R., Roth, J. T., & Fitzpatrick, J. A. (1986, May). Task evaluation form: Development procedures for non-equipment-oriented tasks (AFHRL-TP-85-56). Lowry AFB, CO: Training Systems Division, Air Force Human Resources Laboratory. (AD-A167 411)

APPENDIX A

EXAMPLE TASK EVALUATION FORMS

Examples of completed Forms for three tasks, Building Search, Preflight KC-135A, and Mobility Passport Processing are included in this Appendix. You will notice that there are two different types of pages. The first page of each example (99, 103, and 107) is titled Evaluator Information. The remaining pages of each example are titled Task Evaluation Form.

The Evaluation Information page is used by the evaluator to set up the evaluation and to score observed performance after the evaluation is complete. The Task Evaluation Form pages are used during task performance to guide the evaluator's observations and to record errors. Specific information about the two forms is provided below.

Evaluator Information Page

Blocks A through J provide information to the evaluator about the task to be evaluated and how the task should be presented to the performer for evaluation purposes.

Block K includes a scoring chart which the evaluator completes after the task performance.

Task Evaluation Form Pages

Column 1 lists the steps of the task which will be performed for the evaluation.

Columns 2 through 7 describe what should be evaluated at each step of the task. The evaluator uses this information as standards of task performance. The evaluator circles the corresponding entries on the form when errors occur.

EVALUATOR INFORMATION

A. TASK DESCRIPTION

DATE OF DEVELOPMENT: 6/85 DEVELOPER: Sample

AFSC/DUTY POSITION: Security Police, lead member search team

TASK TITLE: Building Search (20223, no hostages, one armed suspect)

TASK BEGINNING: As performer begins observation of building.

TASK END: When 1 floor has been searched and secured.

STEPS OR EVENTS NOT INCLUDED IN THE EVALUATION: Steps 1 through 6.

B. TASK INFORMATION SOURCES

TITLE	DATE
Education Subject Block Index 0-10	1 May 1981

C. EVALUATION METHOD

Rigged Task Scenario

D. PREVENTATIVE ENVIRONMENTAL CONDITIONS

N/A

E. TASK PRESENTATION

Additional members of search team available.

Building available for use, arrangements coordinated with appropriate BASE personnel.

F. PRESENTATION OF PERFORMER TOOLS, EQUIPMENT, AND MATERIAL

TOOLS, EQUIPMENT, MATERIAL	PRESENTATION
Flashlight	Preselected
M-16	Preselected
Radio	Preselected

G. EVALUATOR TOOLS, EQUIPMENT, AND MATERIAL

Clipboard
Pencil

H. HELP PERMITTED

Other members of team will remain at stairways when directed.

I. EVALUATOR TIME ESTIMATES

Time to Set-Up: 30 minutes

Time to Evaluate: 30 minutes

Time to Reset N/A

J. NUMBER OF EVALUATORS

1

K SCORING CRITERIA

NOTE TO EVALUATOR: It is important that all evaluators score the TEF in the same way. If you have never scored a TEF or are unsure about the number of errors to enter, please see the TEF EVALUATOR INSTRUCTIONS.

EVALUATION AREA	TOTAL POINTS POSSIBLE	NUMBER ASTERISKED ERRORS†	NUMBER NON-ASTERISKED ERRORS		POINTS		POINTS SUBTRACTED
Time/Speed	N/A	N/A	N/A	x	N/A	=	N/A
Sequence-Following	25			x	1.2	=	
End-Product	38		N/A	x	N/A	=	
Safety	25			x	.36	=	
Tools, Equipment, and Materials Use	13			x	.27	=	

Evaluator Comments:

†Circle fail with a score of zero if any asterisked errors occurred.

100 − □ = □

Total Points Subtracted Score

Pass/Fail
Pass = 75-100
Fail = 0-74

TASK EVALUATION FORM

Performer ______________

AFSC ______________

Building Search
Task Title (20223, one armed suspect, no hostages)

Date of Development 6/85

Evaluator ______________

Date ______________

Page 1 of 3

STEP	TIME/SPEED		SEQUENCE-FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequence	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
1 Notify all posted patrols of emergency.	Since evaluation of task performance begins at Step 7. Time is not applicable to this task.					
2 Dispatch patrols.						
3 Designate on scene commander.						
4 Establish command post.						
5 Set-up patrols at strategic locations.						
6 Seal access roads.						
7 Observation of building.			Step 7 through 21		Use movement techniques/ Observe noise discipline/ Observe light discipline/ *Rapid check for suspect activity/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
8 Go through Building 20222.					Use movement techniques/ Observe noise discipline/ Observe light discipline/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
9 Observation.					Use movement techniques/ Observe noise discipline/ Observe light discipline/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
10 Go across landing between buildings.					*Use movement techniques/ Observe noise discipline/ Observe light discipline/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/

TASK EVALUATION FORM

Performer ____________________

AFSC ____________________

Building Search
Task Title (20223, one armed suspect, no hostages)

Date of Development ____________________

Evaluator ____________________

Date ____________________

Page 2 of 3

STEP	TIME/SPEED		SEQUENCE FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
11 Go oast windows.					•Move with low profile/ Observe noise discipline/ Observe light discipline/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
12 Observation of entry window.					Observe noise discipline/ Observe light discipline/ •Rapid check for suspect activity/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
13 Open entry window.					Observe noise discipline/ Observe light discipline/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
14 Observation of entry room.					Observe noise discipline/ Observe light discipline/ •Rapid check for suspect activity/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
15 Search entry room.			•	•Entry room searched and secured.	Enter with low profile/ Observe noise discipline/ Observe light discipline/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
16 Secure N end of 2 floor.			•	•N end of 2 floor searched and secured.	•Use movement techniques/ Observe noise discipline/ Observe light discipline/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
17 Secure N stairway.			•	•N stairway searched and secured.	•Use movement techniques/ Observe noise discipline/ Observe light discpline/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
18 Secure S end of 2 floor.			•	•S end of 2 floor searched and secured.	•Use movement techniques/ Observe noise discipline/ Observe light discipline/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
19 Secure S stairway.			•	•S stairway searched and secured.	•Use movement techniques/ Observe noise techniques/ Observe light techniques/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
20 Search and clear office to right of S stairs.			•	•Office searched and secured.	•Use movement techniques/ Observe noise discipline/ Observe light discipline/ •Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/

TASK EVALUATION FORM

Performer ____________________

AFSC ____________________

Task Title Building Search (20223, one armed suspect, no hostages)

Date of Development 6/85

Evaluator ____________________

Date ____________________

Page 3 of 3

STEP	TIME/SPEED		SEQUENCE-FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
21 Room to room search of 1 floor.			↓	¶ floor searched and secured.	Use movement techniques/ Observe noise discipline/ Observe light discipline/ Use search techniques/	Flashlight, standard, hold out and away from body/ M-16 port arms/ Radio, portable, low volume/
□						
□						
□						
□						
□						
□						
□						
□						
□						

EVALUATOR INFORMATION

A. TASK DESCRIPTION

DATE OF DEVELOPMENT: 6/85

DEVELOPER

AFSC/DUTY POSITION: 431X2

TASK TITLE: Preflight KC-135A

TASK BEGINNING: Beginning of Work Card 1-003

TASK END: End of Work Card 1-010

STEPS OR EVENTS NOT INCLUDED IN THE EVALUATION: Work Card 08 discrepancies found and servicing necessary will only be annotated (i.e., LOX Servicing, Fire Extinguisher Servicing)

B. TASK INFORMATION SOURCES

TITLE	DATE
Technical Orders K-135(K) A-2-9J6-3	3/82
OJT Instructor's Guide No. 8101	1/82

C. EVALUATION METHOD

Actual Equipment Job Environment

D. PREVENTATIVE ENVIRONMENTAL CONDITIONS

N/A

E. PRESENTATION OF OPERATIONAL EQUIPMENT/TASK

No other maintenance being performed

Aircraft requires preflight inspection

Aircraft grounded

F. PRESENTATION OF PERFORMER TOOLS, EQUIPMENT, AND MATERIAL

TOOLS, EQUIPMENT, MATERIAL	PRESENTATION
Minikit	Preselected
Tire Gage	Preselected
Tape Measure	Preselected
Work Cards	Preselected
Rags	Preselected
B-1, B-4, B-5 stands	Preselected

G. EVALUATOR TOOLS, EQUIPMENT, AND MATERIAL

Tire Gage
Work Cards

H. HELP PERMITTED

I. EVALUATOR TIME ESTIMATES

Time to Set Up: N/A

Time to Evaluate: 45 minutes

Time to Reset: N/A

J. NUMBER OF EVALUATORS

1

K. SCORING CRITERIA

NOTE TO EVALUATOR. It is important that all evaluators score the TEF in the same way. If you have never scored a TEF or are unsure about the number of errors to enter, please see the TEF EVALUATOR INSTRUCTIONS.

EVALUATION AREA	TOTAL POINTS POSSIBLE	NUMBER ASTERISKED ERRORS[†]	NUMBER NON ASTERISKED ERRORS		POINTS		POINTS SUBTRACTED
Time/Speed	N/A	N/A	N/A	x	N/A	=	N/A
Sequence Following	10			x	1	=	
End Product	40			x	1	=	
Safety	30			x	6	=	
Tools, Equipment and Materials Use	20			x	5	=	

Evaluator Comments

[†] Circle fail with a score of zero if any asterisked errors occurred.

100 − □ = □

Total Points Subtracted

Score

Pass/Fail
Pass = 75-100
Fail = 0-74

TASK EVALUATION FORM

Performer ____________

AFSC ____________

Task Title Preflight KC-135A

Date of Development Sample

Evaluator ____________

Date ____________

Page 1 of 3

STEP	TIME/SPEED		SEQUENCE FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
1 Test fuel quantity gages.	N/A		1 before 23.	Fuel quantity gages within safe limit.	Ensure correct center of gravity.	
2 Check pitot tubes, check angle of attack, check static boom tubes.				Pitot tubes, ang' .f attack, static boom tubes operational/ not clogged/	Check pitot tube heat with back of hand.	
3 Check rendevous beacon lights for operation.				Beacon lights operational	Lock wheels of stand with all available brakes and install platform pins.	Stand, B-4, lock wheels with all available brakes, install platform pins.
4 Check navigation lights for operation.				Navigation lights operational		
5 Check LOX quantity gage reading.				LOX quantity gage reads at least 6 liters.		
6 Depressurize and check hydraulic quantity gage.				Hydraulic quantity gage reads at least 4.5 gallons.		
7 Close circuit breakers: engine fuel flow and EPR indicators.				Engine fuel flow C.B. and EPR C.B. closed.		
8 Check boom nozzle light bulb for operation.				Boom nozzle light bulb operational		
9 Check portable fire extinguisher.				Fire extinguisher properly secured.		
10 Check foward chinning bar safety pin installed.				Forward chinning bar pin installed.		

TASK EVALUATION FORM

Performer ______________________ Task Title Preflight KC-135A ______________________ Evaluator ______________________

AFSC ______________________ Date of Development Sample ______________________ Date ______________________

Page 2 of 3

STEP	TIME/SPEED		SEQUENCE FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
11 Check portable oxygen cylinders for operation, pressure, and stowed.				Portable oxygen cylinders regulators operational/ stowed/ pressure 280 psi minimum/	Hands clean of oil and grease.	
12 Check nose gear inspection window.				Nose gear inspection window clean.		
13 Check taxi lights.				Taxi lights operational.		
14 Check landing lights.				Landing lights operational.		
15 Check pilot and co-pilot windows.				Pilot and co-pilot windows clean.	Lock wheels of stand with all available brakes and install platform pins.	Stand, B-4, lock wheels with all available brakes, install platform pins.
16 Cargo Compartment: Check portable oxygen cylinders for operation, pressure, and stowed.				Portable oxygen cylinders regulators operational/ stowed/ pressure 280 psi minimum/	Hands clear of oil and grease.	
17 Check portable fire extinguisher properly secured.				Portable fire extinguisher stowed/ properly serviced/		
18 Check gaseous oxygen system serviced.				Gaseous oxygen 400 - 425 psi.		
19 Check main gear and flap emergency crank stowed.				Main gear and flap emergency crank stowed.		
20 Check main gear inspection window.				Main gear inspection window clean		

TASK EVALUATION FORM

Performer ________________

AFSC ________________

Task Title Preflight KC-135A

Date of Development Sample

Evaluator ________________

Date ________________

Page 3 of 3

STEP	TIME/SPEED		SEQUENCE FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type Use
21 Check escape spoiler air supply bottle properly serviced.				Escape spoiler air support properly serviced.		
22 Check AIMS static system drained of moisture.				AIMS static system no moisture in trap.		
23 Check nose gear strut for proper extension				Nose gear strut propr extension.		Tape measure, standard, measure distance between gland nut and strut bottom.
24 Check nose gear tires.				Nose gear tires proper inflation/ no evidence of wear and damage/		Tire gage, standard, insert tape gage onto valve stem and push down smoothly.
25 Check static ports for obstruction.				Static ports clean and unobstructed.		
26 Visually inspect lower fuselage for fuel or hydraulic leaks.				Lower fuselage no fuel or hydraulic leaks.		
27 Inspect fuselage for accumulation of water or fuel.				Fuselage no accumulation of water or fuel beyond allowed limit.		

EVALUATOR INFORMATION

A. TASK DESCRIPTION

DATE OF DEVELOPMENT: 6/85

DEVELOPER: Sample

AFSC/DUTY POSITION: Personnel Outbound

TASK TITLE: Mobility Passport Processing

TASK BEGINNING: When member brings completed forms to office.

TASK END: When DD Form 1056 is suspensed.

STEPS OR EVENTS NOT INCLUDED IN THE EVALUATION: Member completes DSP 11 and Form 1056.

B. TASK INFORMATION SOURCES

TITLE	DATE
AFR 30-4	

C. EVALUATION METHOD

Actual Job Enviroment

D. PREVENTATIVE ENVIRONMENTAL CONDITIONS

N/A

E. TASK PRESENTATION

Member brings in completed Form 1056, DSP 11, and acceptable proof of citizenship.

Photos are available.

Evaluate when Mobility Passport Processing is performed.

F. PRESENTATION OF PERFORMER TOOLS, EQUIPMENT, AND MATERIAL

TOOLS, EQUIPMENT, MATERIAL	PRESENTATION
AFR 30-4	Performer selects

G. EVALUATOR TOOLS, EQUIPMENT, AND MATERIAL

AFR 30-4
Clipboard
Pencil

H. HELP PERMITTED

Performer can refer to regulations at any time.

I. EVALUATOR TIME ESTIMATES

Time to Set-Up: None

Time to Evaluate: 20 minutes

Time to Reset: None

J. NUMBER OF EVALUATORS

1

K. SCORING CRITERIA

NOTE TO EVALUATOR: It is important that all evaluators score the TEF in the same way. If you have never scored a TEF or are unsure about the number of errors to enter, please see the TEF EVALUATOR INSTRUCTIONS.

EVALUATION AREA	TOTAL POINTS POSSIBLE	NUMBER ASTERISKED ERRORS[1]	NUMBER NON ASTERISKED ERRORS		POINTS		POINTS SUBTRACTED
Time/Speed	N/A	N/A	N/A	x	N/A	=	N/A
Sequence-Following	17			x	17	=	
End-Product	50			x	3.1	=	
Safety	N/A	N/A	N/A	x	N/A	=	N/A
Tools, Equipment, and Materials Use	33			x	11	=	

Evaluator Comments.

[1] Circle fail with a score of zero if any asterisked errors occurred.

100 − □ Total Points Subtracted = □ Score

Pass/Fail
Pass = 75-100
Fail = 0-74

TASK EVALUATION FORM

Performer ____________________

AFSC ____________________

Task Title Mobility Passport Processing

Date of Development 6/85

Evaluator ____________________

Date ____________________

Page 1 of 2

STEP	TIME/SPEED		SEQUENCE-FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
1 Review completed DSP 11.				*DSP 11 complete.		
2 Review completed DD Form 1056				*DD 1056 complete.		
3 Review proof of citizenship; accept or reject.				*Proof acceptable.		Regulation 30-4 determine if proof is acceptable.
4 Enter info. from ID card on DSP 11.						
5 Ensure individual signs back of pictures.				Pictures signed on reverse side.		
6 Affix one picture to the DSP 11.				Photo affixed in space provided.		
7 Swear individual in.			Step 7 before 8.	*Individual sworn in. ✓		*Regulation 30-4 swear individual in.
8 Ensure individual signs DSP 11.				DSP 11 signed.		
9 Attach DSP 11, second picture, proof of citizenship, DD Form 1056.				*Package complete including DSP 11, second picture, proof of citizenship, DD Form 1056.		
10 Mail package to State Department.				*Package sent to appropriate department.		

TASK EVALUATION FORM

Performer ______________

AFSC ______________

Task Title Mobility Passport Processing

Date of Development 6/85

Evaluator ______________

Date ______________

Page 2 of 2

STEP	TIME/SPEED		SEQUENCE-FOLLOWING	END-PRODUCTS	SAFETY	TOOLS, EQUIPMENT, AND MATERIALS USE
1. Description	2. Start/Stop Measuring	3. Standard	4. Sequences	5. End-Product Criteria	6. Procedures and Regulations	7. Item, Size or Type, Use
1 Suspense DD Form 1056 for 75 days.				Form 1056 suspensed for 75 days.		
□						
□						
□						
□						
□						
□						
□						
□						
□						

APPENDIX B

BLANK WORKSHEETS

Worksheet 01: Task Steps

Task: Developer:

Column A: #	Column B: Step Description

Worksheet 02: Task Definition

Task: Developer:

Line A AFSC/Duty Position:

Line B Task Title:

Line C Task Beginning:

Line D Task End:

Block E Steps or Events not Included in the Evaluation

Block F Task Information Sources

Title	Date

Worksheet 03: Evaluation of Time or Speed of Task Performance

Task:

Developer:

Column A Critical Segments	Block B Starting Point	Block C Stopping Point	Block D Standard

Worksheet 04: Evaluation of Sequence-Following

Task: Developer:

Block A Series of Steps

Step ____ through ____

Step ____ through ____

Step ____ through ____

Step ____ through ____

Step ____ through ____

Block B Single Steps

Step ____ before Step(s) ______________________________

Step ____ before Step(s) ______________________________

Step ____ before Step(s) ______________________________

Step ____ before Step(s) ______________________________

Step ____ before Step(s) ______________________________

Worksheet 05: Evaluation of End-Product

Task:

Developer:

Column A End-Products	Column B Criteria	Column C Step

Worksheet 06: Evaluation of Safety Procedures and Regulations

Task:

Developer:

Column A Safety Procedures and Regulations	Column B Steps

Worksheet 07: Evaluation of Tools, Equipment, and Materials Use

Task:

Developer:

Column A Tools, Equipment and Materials Use	Column B Size/Type	Column C Correct Use	Column D Steps

Worksheet 08: Evaluation Scenario

Task:

Developer:

Line A Evaluation Method:

Block B Preventative Environmental Conditions

Block C Task Presentation

Block D Help Permitted

Block E Presentation of Performer Tools, Equipment, and Materials

Tools, Equipment, and Materials	Presentation

Block F Evaluator Equipment

Block G Evaluator Time Estimates

Time to set up ____________
Time to evaluate ____________
Time to reset ________

Block H Number of Evaluators ____________

Worksheet 09: Scoring Criteria

Task: ______________________ Developer: ______________________

Time/Speed

Rank □ A

All Entries □ B + Asterisked Entries □ C = Enter Total in Block E

Points For Time/Speed From Appendix D □ D ÷ Total B + C □ E = Points Per Non Asterisked Entry □ F

Sequence-Following

Rank □ A

All Entries □ B + Asterisked Entries □ C = Enter Total in Block E

Points For Sequence-Follow ng From Appendix D □ D ÷ Total B + C □ E = Points Per Non-Asterisked Entry □ F

End-Product

Rank □ A

All Entries □ B + Asterisked Entries □ C = Enter Total in Block E

Points For End-Product From Appendix D □ D ÷ Total B + C □ E = Points Per Non-Asterisked Entry □ F

Safety

Rank □ A

All Entries □ B + Asterisked Entries □ C = Enter Total in Block E

Points For Safety From Appendix D □ D ÷ Total B + C □ E = Points Per Non-Asterisked Entry □ F

TEM Use

Rank □ A

All Entries □ B + Asterisked Entries □ C = Enter Total in Block E

Points For TEM Use From Appendix D □ D ÷ Total B + C □ E = Points Per Non-Asterisked Entry □ F

APPENDIX C

RATING FORMS

INSTRUCTION SHEET

Rater No. ______________ Name_____________________

You will use the items included on these Rating Sheets to rate the information listed on a Task Evaluation Form. There are three groups of questions referring to the (1) Accuracy; (2) Completeness; and (3) Criticality of the information listed on the Task Evaluation Form. There are six questions in each group. One question refers to each of the five Performance Measures and one question refers to the Task Evaluation form as a whole.

You will rate the information on the Task Evaluation Form using a rating scale of 1 to 7. You should indicate your rating for each question by circling the appropriate number.

You are to rate the forms by yourself, without discussion with other panel members. It is important that you rate the Task Evaluation Forms in the order in which they have been presented to you. The Rating Sheets each have a form number in the upper right corner which should correspond to the number on the Task Evaluation Form being rated. First, rate all forms for Accuracy, then for Completeness, and lastly rate all forms for Criticality. Make sure you understand the distinctions between these categories before you rate any of the forms.

Once you have rated a Task Evaluation Form, do not return to it to alter already-made responses, return only to complete any blanks that remain.

Accuracy

To rate the accuracy of the information on the Task Evaluation Form, you should consider the information listed in Columns 2 through 7 (the appropriate column numbers are noted next to each question). Ask yourself:

TO WHAT EXTENT IS THE INFORMATION LISTED ON THE TASK EVALUATION FORM ACCURATE AND CONSISTENT WITH AIR FORCE POLICY AND TASK DOCUMENTATION?

1 = The information is totally inaccurate and inconsistent with Air Force policy and task documentation.

7 = The information is totally accurate and consistent with Air Force policy and task documentation.

1.	Time	Columns 2-3	1	2	3	4	5	6	7
2.	Sequence-Following	Column 4	1	2	3	4	5	6	7
3.	End Product	Column 5	1	2	3	4	5	6	7
4.	Safety	Column 6	1	2	3	4	5	6	7
5.	Tools, Equipment, and Materials	Column 7	1	2	3	4	5	6	7

Criticality

To rate the criticality of the information on the Task Evaluation Form, you should consider the steps in Column 1 which have been identified for the Performance Measures. Ask yourself:

TO WHAT EXTENT ARE THE STEPS IDENTIFIED ON THE TASK EVALUATION FORM CRITICAL TO SUCCESSFUL TASK PERFORMANCE?

1 = The steps are not critical; none of the steps identified are critical to successful task performance.

7 = The steps are critical; all f the steps identified are critical to successful task performance.

6. Time 1 2 3 4 5 6 7

7. Sequence-Following 1 2 3 4 5 6 7

8. End Product 1 2 3 4 5 6 7

9. Safety 1 2 3 4 5 6 7

10. Tools, Equipment, and Materials 1 2 3 4 5 6 7

Completeness

To rate the completeness of the information on the Task Evaluation Form, you should consider the steps in Column 1 which have been identified for the Performance Measures. Ask yourself:

TO WHAT EXTENT ARE THE STEPS IDENTIFIED ON THE TASK EVALUATION FORM COMPLETE?

1 = The steps are not complete; none of the steps necessary to evaluate the task performance have been identified.

7 = The steps are complete; all of the steps necessary to evaluate task performance have been identified.

11. Time	1	2	3	4	5	6	7
12. Sequence-Following	1	2	3	4	5	6	7
13. End Product	1	2	3	4	5	6	7
14. Safety	1	2	3	4	5	6	7
15. Tools, Equipment, and Materials	1	2	3	4	5	6	7

☆U S GOVERNMENT PRINTING OFFICE 1986 - 659 - 055/40026