Navy Personnel Research and Development Center
San Diego, California 92152-6800 TN-91-11 April 1991

Troubleshooting Assessment and Enhancement (TAE) Program: Theoretical, Methodological, Test, and Evaluation Issues

Harry B. Conner
Frank H. Hassebrock

NPRDC-TN-91-11 April 1991

Troubleshooting Assessment and Enhancement (TAE) Program: Theoretical, Methodological, Test, and Evaluation Issues

Harry B. Conner
Frank H. Hassebrock

Approved and released by
J. C. McLachlan
Director, Training Systems Department

Approved for public release;
distribution is unlimited.

Navy Personnel Research and Development Center
San Diego, California 92152-6800

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATE COVERED
	April 1991	Technical Note--Oct 87-Mar 90

4. TITLE AND SUBTITLE
Troubleshooting Assessment and Enhancement (TAE) Project: Theoretical, Methodological, Test and Evaluation Issues

5. FUNDING NUMBERS
0603720N-R1772-ET01

6. AUTHOR(S)
Harry B. Conner, Frank H. Hassebrock

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Navy Personnel Research and Development Center
San Diego, California 92152-6800

8. PERFORMING ORGANIZATION REPORT NUMBER
NPRDC-TN-91-11

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Chief of Naval Operations (OP-11), Navy Department,
Washington, DC 20350-2000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
For additional information, see Rep. Nos. NPRDC-TN-91-12 and NPRDC-TN-91-13.

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

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

The purpose of the Troubleshooting Assessment and Enhancement (TAE), R&D effort was to develop, test, and evaluate a computerized system to provide the Navy with a troubleshooting assessment and training capability. This technical note presents the results of the literature review, the theoretical and methodological issues that were to be considered, and the proposed test and evaluation plan for the TAE effort.

14. SUBJECT TERMS
Troubleshooting, simulation, performance evaluation, troubleshooting training

15. NUMBER OF PAGES
91

16. PRICE CODE

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

FOREWORD

The Troubleshooting Assessment and Enhancement (TAE) Program previously titled Troubleshooting Proficiency Evaluation Program, TPEP), was sponsored by the Deputy Chief of Naval Operations (OP-11) and was performed under 0603720N-R1772-ET01. The purpose of the TAE program was to develop a low-cost, microcomputer-based system to provide an objective measure of the troubleshooting proficiency of Navy technicians.

Three technical notes document the TAE program. This technical note, which is the first in the series, presents the results of the literature search, the methodology for developing a troubleshooting proficiency evaluation system, and the resulting test and evaluation plan. The second technical note presents the design and development of the computerized troubleshooting proficiency evaluation system (Conner, Poirier, Ulrich, & Bridges, 1991). The final technical note presents the results of the test and evaluation as well as the conclusions and recommendations for enhancing the TAE delivery system (Conner, Hartley, & Mark, 1991).

The authors wish to acknowledge the assistance of David Dickinson and Sandra Hartley of Instructional Science and Development, Inc. in the preparation of the Test and Evaluation Plan (Appendix B).

J. C. McLACHLAN
Director, Training Systems Department

SUMMARY

Problem

The Navy requires improved methods for assessing the troubleshooting proficiency of its technicians. There was no consistent way to assess personnel performance or the transfer of training in this skill area to the operational environment.

Purpose

The purpose of the Troubleshooting Assessment and Enhancement (TAE) Program was to develop a low-cost microcomputer-based system to provide an objective measure of troubleshooting proficiency in support of Navy technicians. This technical note presents the results of the literature review, the theoretical and methodological issues considered during design and development, and the proposed test and evaluation plan.

Approach

The approach of the TAE program was to select hardware and test sites; develop operational procedures; define TAE technology; develop application methods; select a demonstration delivery system; select troubleshooting tasks; and evaluate and develop troubleshooting scenarios. Once the delivery system and scenarios were developed, the test and evaluation (T&E) phase would be accomplished.

This technical note provides an overview of research objectives for the measurement and evaluation of troubleshooting proficiency; presents the traditional measures of job proficiency which were reviewed; provides additional information on the literature reviewed; and presents the test and evaluation plan developed.

Results

Results of the literature review are presented as related to the effort to empirically determine the appropriate measures and evaluation approach to be used in TAE research, development, test, and evaluation. Measurement and evaluation techniques, particularly to ensure validity and reliability are examined. The behavioral and cognitive task analysis approaches are discussed. An attempt to integrate the literature relating to these analysis approaches and TAE are provided.

Methods for predicting job-ability are presented and the characteristics that relate to troubleshooting (TS) ability and the impact on the procedures necessary to develop a TS skill assessment capability are discussed. A three-step process for developing a TAE system is described: describe troubleshooting characteristics; operationally define the constructs and develop measures: and specify the methodology for developing composite score of troubleshooting proficiency.

Recommendations for TAE Development

1. TS scenarios should be developed by expert technicians.

2. Scenarios should be tested (verified) by a different group of experts.

3. Outcome and process variables should be measured and analyzed separately.

CONTENTS

INTRODUCTION

Problem

The Troubleshooting Assessment and Enhancement (TAE) program supports Fleet and Manpower, Personnel, and Training (MP&T) community requirements. These requirements include reducing mean time to repair (MTTR), increasing mean time between failures (MTBF), reducing no fault removals (NFRs) while providing fleet evaluation/training, assisting in the accomplishment of the Office of the Chief of Naval Operations (OPNAV) requirement of training assessment and feedback, and contributing to fleet readiness via on-board-training (OBT) within the context of Total Forces application; i.e., the resultant program will apply and support active duty and reserve forces.

The original hardware test system selected for TAE design, development, test and evaluation was the Naval Modular Automated Communications System (V)/Satellite Communication (NAVMACS/SATCOM) system, which is maintained by the occupational community of Electronics Technicians (ETs) with the Navy Enlisted Classification (NEC) number of ET-1453.

The Navy has a critical need to improve fleet maintenance capability and alleviate maintenance problems through training and aiding technologies (Nauta, 1985). Currently, the Navy has limited means of objectively measuring the troubleshooting proficiency of the shipboard technicians and their ability to contribute to operational readiness. Other than subjective supervisory opinion, there is no consistent way to assess the transfer of training, particularly hands-on, in the Navy "C" (i.e., hardware systems) schools. Once the "C" school graduate has been integrated into the ship's force, fleet commanders have no objective method to assess the technician's performance capabilities or skill degradation over time. In addition, the schools receive little quantifiable feedback identifying specific areas where troubleshooting training requires greater emphasis or improvement.

Due to limited availability of system hardware at the "C" schools, actual hands-on training time is severely restricted. This minimizes the amount of time students explicitly use their system knowledge and, therefore, limits the effectiveness of instructional programs. Once on-board, the ship safety hazards associated with corrective maintenance of weapon system hardware preclude the use of drill and practice exercises. This limits the technicians' ability to maintain their troubleshooting skills and restricts maintenance or improvement of his abilities.

Purpose

The purpose of the Troubleshooting Assessment and Evaluation (TAE) program was to develop a low-cost microcomputer-based system to provide an objective measure of troubleshooting proficiency of Navy technicians. Specifically, the TAE program was to (1) assess personnel troubleshooting capabilities within the Navy training environment (e.g., "C" school and/or reserve training activities), (2) develop drill and practice for personnel in training awaiting hardware availability or active duty assignments, (3) improve curricula and training methods based on school troubleshooting assessment results, (4) provide fleet and SELRES on-board training (OBT) through drill and practice exercises, (5) assess fleet and reserve personnel troubleshooting capabilities, (6) develop an objective measure of operational readiness of fleet and reserve personnel in the area of systems hardware troubleshooting capability, (7) improve operational

readiness, and (8) improve curricula and instructional methods as a result of objective operational fleet and SELRES feedback of assessment/evaluation data to the training community.

This technical note provides the results of the literature search that preceded the TAE effort. It also presents the issues that were addressed during the initial design and development phases and the proposed test and evaluation plans.

Conner, Poirer, Ulrich, and Bridges (1991) present the program and software design, development and administration; and Conner, Hartley, and Mark (1991) present the test and evaluation results.

The approach of the TAE program was to (1) select hardware system and test sites, (2) develop operational procedures at the test sites, (3) define TAE technology, (4) develop applications methodology, (5) select a TAE demonstration delivery system, (6) select troubleshooting tasks, (7) evaluate troubleshooting tasks for scenario development, and (8) develop troubleshooting scenarios.

Once the scenarios/episodes were developed, a troubleshooting factors assessment model was developed, and, using data results from the test sites, compared across and within sites to ensure accuracy of, and improve, the model (Conner, Hartley, & Mark, 1991).

Background

The Troubleshooting Assessment and Enhancement (TAE) Program comprised a related set of research and development efforts that were developed to objectively measure troubleshooting skills. The first effort investigated the high-technology occupational community of electronics maintenance. Originally, the program was developed to provide an objective measure of evaluating troubleshooting (TS) proficiency that could be used to compare the Enlisted Personnel Individualized Career System (EPICS) personnel, and the conventional personnel system (CPS) personnel (Conner 1986, 1987). The TAE subsequently became a research effort because of its (TAE) perceived utility as a method to assess TS capabilities of "C" school students, shipboard technicians, reserve personnel, and general fleet readiness. For all of these assessment groups, TAE was also considered potentially useful in providing TS drill and practice in a variety of settings and problem domains. The combination of assessment and practice should allow for an ongoing test and evaluation process providing relevant information for the improvement of TS curricula, training methods, and assessment techniques.

Overview of Research Objectives

The general research objectives driving the TAE concern the (1) measurement of troubleshooting performance and (2) subsequent discrimination of proficiency levels in troubleshooting. Hence, the project was pursued as a test-and-evaluate endeavor and the results will provide inputs for a variety of other pursuits including training, simulation, and performance assessment. As a general measurement endeavor, any test should provide a reliable and valid means of representing aspects of TS performance. Initially, the program was interested in going beyond simply describing or summarizing TS behavior since measurement was assumed to be a process for representing troubleshooting proficiency. Such an attempt, therefore, involves more

than behavioral measurement since a complex attribute such as TS proficiency is realized through a variety of perceptual, cognitive, and motor behaviors. In other words, TS proficiency is a multifaceted attribute and any measure may be relevant to more than one aspect of the attribute or even to several other related attributes. As a result of this broader perspective, both behavioral and cognitive protocol analyses were planned.

Measures of Job Proficiency

Researchers in the areas of personnel and training have distinguished between different measures of job proficiency (e.g., Harris, Campbell, Osborn, & Boldovici, 1975). Three general types of evaluation instruments used to measure job proficiency include (1) job performance ratings, (2) job knowledge tests, and (3) job performance tests. Job performance ratings have several limitations centering around a lack of standardization and objectivity (e.g., rater errors and biases). Although job knowledge tests have been extensively used, several researchers have found very low correlation between such tests and measures of actual job performance (e.g., Foley, 1977). Job performance tests also have several factors which constrain their use including cost, time, resources, safety, and availability.

Vineberg and Joyner (1982) reviewed reports published between 1952 and 1980 that investigated the prediction of job performance of enlisted personnel in the U.S. military establishment. They also distinguished between measures of job proficiency and measures of job performance. Job proficiency refers to skills and knowledge needed for performing a job while job performance refers to actual job behavior. They found that job proficiency was measured by (1) paper-and-pencil knowledge tests, (2) measures of task performance which simulate complete job tasks, and (3) measures of task element performance which simulate components of tasks. Job performance is typically measured by some form of rating; for example, (1) global, (2) job element, (3) productivity, or (4) grade or skill level. Several studies have found that ratings do not provide consistent discrimination of job performance.

The review of studies shows generally low correlations between job sample tests and job knowledge tests. Most knowledge tests focus on theoretical and terminological information that are not directly descriptive of job performance; that is, you do not need to know theory (electronics engineering) to do the job (repair electronic equipment). However, knowledge tests do show higher correlations with performance if such tests are developed to assess information and behaviors directly related to job performance (Vineberg & Joyner, 1982); i.e., you do need to know the job (know how to do electronic repair things) to be able to do the job (be able to perform the tasks). Correlations between job performance ratings and job sample tests are also low. The correlations are slightly higher between ratings and knowledge tests. Overall, the typical measures of job proficiency and performance show very low relationship. Furthermore, predictor variables (generally demographic and individual data) are most useful for the criterion measure of job knowledge and least useful for global job performance ratings. Aptitude measures (e.g., Armed Services Vocational Aptitude Battery (ASVAB)), generally show very low correlation with job performance criteria. For example, Armed Forces Qualification Test (AFQT) had correlations of about .30 with job sample tests (Vineberg & Taylor, 1972), while experience (months on job) had correlations ranging from.39 to .69. Mackie, McCauley, and O'Hanlon (1978) also found that ASVAB tests had a median correlation of 0.0 with job performance tests.

Vineberg and Joyner (1982) claimed that job performance tests are not necessarily required for many needs of evaluation and measurement since most complex job behaviors are mediated by information. They claim:

> When technical proficiency is a relevant aspect of performance, tests of job knowledge may provide the most objective, practical means for assessing it, despite their general dependence on verbal ability. Even though they have sometimes been found not to correlate well with performance tests or job experience, job knowledge tests can share considerable variance with both if designed from carefully developed job analysis data.

Vineberg and Joyner (1982) further claimed that performance in training is currently the best predictor of job proficiency (measured as job knowledge) and job performance (measured by supervisor ratings). They suggest two approaches to maximize prediction: (1) the use of miniaturized training and assessment centers that will allow the trainee to experience samples of work activities and (2) individualized, self-paced training.

Pickering and Bearden (1984) reviewed studies from 1953 through 1981 that used job performance tests to measure the skills of individuals in the Navy. They, in a somewhat more pragmatic fashion, defined a job performance test as one that "measures job skills by requiring examinees to perform specific tasks, under controlled conditions, in an identical or similar fashion to that required on the actual job." One project investigated troubleshooting skills by requiring subjects to respond to symbolic tasks (named, AUTOMASTS; Bryan, 1954). Researchers used a series of AUTOMASTS troubleshooting problems and a series of paper-and-pencil electronic job knowledge tests as predictors for a criterion of job success. Their goal was to combine the predictor scores into a single, composite score that would provide "the best measure of troubleshooting ability." The machine used to present the problem recorded a sequential list of every check and replacement made, a record of time spent on the problem, and whether the problem was solved. Experts were to examine these measures in order to formulate a best score of overall performance. However, as different experts might recognize different aspects of behavior as desirable, three judges with different experience and background were selected to classify the individual records from 10 problems on the basis of "overall goodness of performance." Each composite judgment was classified into one of five categories of effectiveness. Agreement was evaluated by using analysis of variance (ANOVA) procedures to calculate coefficients of interrater agreement. Correlations had a median of .91 and they concluded the ratings could be used as a "more ultimate" criterion. Twelve AUTOMASTS TS performance measures (number of actions, average time, redundancy, number solved, first replacement effectiveness, component replacement score, direct clue actions, indirect clue actions, errors, clue quality, proximity, and neutral actions) were intercorrelated with the judgment standard. The direct clue actions score was the most highly correlated (.85). The standard score of this measure was combined with the paper-and-pencil test standard score to form a single criterion, named electronics proficiency score (EPS). EPS scores were then correlated with several other variables (e.g., months of experience, school grades, etc.). However, the project was never implemented on a larger scale.

Pickering and Anderson (1976) concluded from their survey of performance measurement research literature that the degree of simulation required is important in developing a job performance test:

It may be necessary to test on the actual operational equipment; an equipment mockup may be required; some form of computerized simulation may be appropriate; a pictorial representation questions may be sufficient and a performance test is not required.

They also concluded that there was a need for further experimental investigations to provide information relative to job-performance test objectivity, reliability, validity, degree of simulation required. and use in testing programs (see also Arima & Neil, 1978; Wetzel, Konoske, & Montague, 1983).

Measurement Validity and Reliability

Traditional approaches to measurement and evaluation depend heavily on the determination of test validity and reliability. Many of the TS tests that have been employed over the years were evaluated on the basis of their content validity. For the most part, studies have constructed achievement tests for measuring TS performance and the major concern was to develop tests with a high degree of content validity; that is, the test items contain a representation of a specified universe of content. Achievement tests have content validity to the extent that they adequately sample the content in a domain. Therefore, the test directly measures performance and there is no need to compare the test results with some other criterion. In summary, tests which need to achieve content validity do not logically need any additional empirical criterion. One approach, therefore, for measuring job performance relies on tests that achieve content validity through adequate sampling of task or job behaviors (e.g., Foley, 1974, 1975, 1977; Shriver & Foley, 1974). The debate over training simulators also involves the issue of content validity regarding fidelity of simulation.

Tests may also include measures intended to establish functional relationships with a particular variable. This type of test requires the determination of empirical validity which may vary between specific temporal relationships; predictive validity suggests that the test will specify performance in some future behavior (or attribute) whereas concurrent validity is concerned with the relationship between the predictor test and any other contemporary measurement. Many researchers have attempted to validate job knowledge tests and job performance ratings by seeking an ultimate empirical criterion such as actual job performance. However, this ultimate criterion does not become available often enough for practical use. This may have to do with the selection of the test groups; that is, definition of expert vs. novice. In any case a near-ultimate or secondary criterion must be selected and these typically have taken form as some type of job performance task or job component performance task (e.g., job test performance tests (JTPTs) as in Foley, 1974). However, the use of JTPTs is not without debate as to their effectiveness in training or selection (Pickering & Bearden, 1984). The most serious limitations in studies of JTPT have been their omission of assessment of validity and reliability. Only 4 of 34 studies reported information on concurrent validity and 2 contained information on content validity. Furthermore, the tests often failed to discriminate between experienced and inexperienced groups (Pickering & Anderson, 1976). Even more discouraging is the lack of correlation between symbolic performance tests (i.e., paper and pencil tests) and either JTPT or actual job performance (Foley, 1974). Alluisi (1977) stressed that the criterion problem in performance assessment is critical for many applications including the validation of selection and training techniques, improvements in man-machine systems, and the establishment of optimum operator demands in task performance.

A review of job performance testing research conducted over the past 30 years at NPRDC suggests that the Navy does not have a comprehensive system for measuring the job performance capabilities of individuals (Pickering & Bearden, 1984). The limitations of measuring job performance suggest similar problems may occur in measuring TS proficiency. The shortcomings include:

1. Evaluations tended to stress general capabilities instead of specific performance deficiencies.

2. Evaluations were typically based on paper-and-pencil tests or supervisor's ratings instead of job-performance tests.

3. Some of the studies tried to test an entire population instead of using sampling procedures and were consequently overwhelmed by the demands of the effort.

4. Many evaluations were conducted aboard ship where there were many constraining factors.

Measurement of Electronic Troubleshooting Proficiency

Nauta (1985) reviewed the impact of maintenance training on fleet maintenance problems and formed a research question similar to the ones asked by others, "What distinguishes exactly an expert troubleshooter from an average technician?" Over 30 years of research on TS has not led to any agreement on the nature of TS proficiency, let alone its measurement. However, Nauta also suggests that one problem within this dilemma is that traditional training and instructional programs (and hence, assessment and evaluation) have relied on tasks analyses to identify behavioral objectives that are appropriate for proceduralized tasks but are not appropriate for specification of cognitive skills underlying complex performance as in TS. Nauta proposes that:

> Development of troubleshooting skills requires more than learning a few standard rules and applying them in free-play or operational equipment; it requires a thorough understanding of theory of operations, functional interdependencies, and symptom-cause relationships in order to form a cognitive map or model of the system to be maintained.

Once again this line of reasoning seems to dictate that any assessment or measurement of TS proficiency demands both empirical (i.e., predictive or concurrent) validity and construct validity.

Information gathered from structured interviews with eight experts in electronics maintenance suggested that the Navy believes a persistent myth about the true difficulty of the tasks involved in maintenance. The Navy appears to believe that the task is simpler than it really is and systematically underestimates the amount of training and experience required for successful performance (Parker & Dick, 1985). According to these experts concerning measurement of skill proficiency, "At the present time, there does not appear to be an effective, yet simple, way to measure technician proficiency." These experts suggest that experiments be conducted to compare the performance of highly skilled technicians with those of lesser ability: "The objective of such a comparison would be to identify the critical skill and knowledge elements that permit experienced technicians to excel at their jobs." They suggest a collection of performance measures that include traditional TS product measures (e.g., accuracy and time) but also measures that reflect current

understanding of the perceptual and cognitive skills necessary for complex TS performance. One implication of this suggestion is that there may not be a simple way to assess TS proficiency if studies only rely on the traditional approach adopted in previous studies.

TAE was designed to provide a standard method of assessing TS proficiency across a number of different fault scenarios. Obviously, a major issue to be addressed concerns the use of TAE for discriminating between levels of TS proficiency. Since TAE represents a testing (i.e., classification) or measurement procedure, it is necessary to consider requirements of validity and reliability. The microcomputer-delivered testing context affords a great deal of face validity regarding TS tasks (Conner 1986,1987). For the purposes of TAE, a test of TS proficiency could be evaluated on the basis of its predictive validity in the sense that student performance in "C" school should document readiness for fleet TS. Unfortunately, it is not clear if there is any established criterion of fleet TS (unless it would be shipboard TS under real conditions; i.e., nonroutine, nonplanned preventative or training situations). Criterion validity may also be addressed by determining the degree to which TAE discriminates TS ability through correlations with other independent measures of TS proficiency. The research literature suggests several individual abilities (context free) and behaviors (either context-free or task-specific) that are related to TS.

The problems of achieving criterion validity in developing and using job performance tests (e.g., JTPT) may be indicative of a need for test developers to address issues of construct validity. Troubleshooting proficiency embraces a multifaceted collection of perceptual, cognitive, and motor behaviors; therefore, TS should be viewed as a construct defined by a large number of related skills and abilities (cf. Parker & Dick, 1985). Many of the TS studies have only been concerned with observable, easily quantifiable behaviors (e.g., success or failure, time) of TS performance. Furthermore, since the TAE project also embraces instructional, training, and general evaluation goals, a successful determination of criterion validity will serve as an input to additional determination of construct validity. This requirement necessitates that any test and evaluation (T&E) effort will, ultimately, need to ensure that data collection and analysis attend to conceptual clarification of cognitive, as well as behavioral, indicators of TS knowledge, skill, and ability. Clarification of the construct validity of TAE will therefore assure that measurement and instructional goals will be properly served.

LITERATURE REVIEW

Background

Glaser (1976) has outlined the components of instruction that provide a prescriptive set of design procedures to be mapped onto by descriptive, substantive concepts of human learning. The design of instructional (training) procedures involve the following components:

1. Analysis of competence (i.e., the state of knowledge and skill that is to be achieved).

2. Description of the initial state of the learner.

3. Conditions and procedures that can be used to bring about a transformation from the initial state to the desired state of competence.

4. Assessment and evaluation of the outcomes provided by the instructional conditions.

The analysis of competence essentially answers the question, "What is to be learned?" However, since the instructional materials and procedures of the Navy schools have already been developed, this first design component of TAE addresses a slightly different question "How does a competent troubleshooter perform that distinguishes him from less competent troubleshooters?" In this context, the analysis of competence will identify properties associated with skillful and less skillful TS performance.

Traditionally, a task analysis provides an analytic description of properties underlying a skill and produce a specification of the behavioral objectives which define a criterion performance. A task analysis can also provide a description of performance (or skill) in terms of the demands upon the learner's cognitive processes. A task analysis was planned in TAE after completion of the efforts to design, develop, test and evaluate a process that would provide discriminatory and behavioral data that would lead to a behavioral protocol or troubleshooting model. Once the behavioral protocols were accomplished, cognitive protocols could be pursued. Although the cognitive task analysis was not conducted, the proposed approach included in this TN. Cognitive task analyses can specify (1) the knowledge structures required for performance (issues of knowledge content, representation, and organization) and (2) the cognitive processes or procedures (i.e., heuristics, strategies, problem representations) which access and use such knowledge. The task analysis specifies the behaviors, knowledge, and cognitive procedures that underlie a specific level of performance; in other words, these analytic descriptions become the targets of instruction and training. In essence, assessment of TS proficiency (not to mention instructional and training design) must be based on the demand for criterion and construct validity.

Given the above arguments, this literature review integrates the existing literature on TS (based heavily on rational and behavioral task analyses) with the developing literature on the cognitive skills which distinguish levels of expertise in specific problem domains (based on experimental studies reflecting a cognitive task analysis). The integration of these approaches was used in the design of the TAE system. The operational TAE system is reported in Conner, Hartley, and Mark (1991).

Development Procedures

Rose, Fingerman, Wheaton, Eisner, and Kramer (1974) have proposed methods for predicting job-ability requirements. Although TAE is not being evaluated as a selection instrument, its goal of performance assessment requires many of the same procedural steps. First, a task or job analysis must be done in order to form hypotheses as to what behavioral characteristics distinguish levels of successful performance. Characteristics will be identified that will be related to degrees of TS proficiency. These hypotheses are then translated into a definition of the troubleshooter characteristics necessary for performance. At this point, the previous literature reviews in TS and related studies of problem solving in other professional domains provide a catalog of the psychological constructs comprising TS behavior (including perceptual, cognitive, and motor components).

Second, the constructs are operationally defined through the development of a set of possibly useful tests (i.e., tasks) that will provide a relevant set of measures of these characteristics. These

tests and measurements should be evaluated for both reliability and validity. Validity may be divided into empirical validity (i.e., concurrent), which would involve a criterion measure, or construct validity, which should differentiate among groups of technicians differing in TS skill. These steps embrace Glaser's suggestions (1976) in that a task analysis must first identify components of competence (at various levels) in order for the researcher to develop relevant and appropriate measures for assessing performance proficiency.

Third, one goal of the TAE evaluation is to determine, which criteria should be used to discriminate among levels of TS proficiency. This goal is constrained by the same criterion problem that has characterized the past three decades of TS research, especially in the areas of job performance and job proficiency tests.Two possible solutions to this dilemma come to mind. One approach will concern the determination of a scheme to discover a valid and reliable composite score of troubleshooting proficiency.

The other solution based on Glaser (1976) and Rose et al. (1974) to this dilemma may be to realize an effort toward specification of the constructs underlying TS performance in addition to attempts to ensure criterion validity. To continue this line of reasoning, therefore, is to realize that the search for the ultimate criterion (Foley, 1977) may be a constraint, especially in the cases where the criterion for prediction is simply a pass or fail discrimination. Mallory and Elliott (1978) in their review of studies that have used simulations to assess TS performance assert that "even if the criterion test produces valid and reliable measures of performance, it may still fail in usefulness because it is not diagnostic." If the criterion is only pass or fail, "such a criterion would not be useful if we were interested in knowing why those who failed did so, performance aids would be required to bring them to an acceptable level of competence." Again, this type of limitation has its origins in the reliance on simple behavioral products of TS performance such as success or failure or gross measures of behavior. Studies which rely on quantitative analyses of such product measures are perfect examples of the problems which result from the fact that "a diversity of behavior may be hidden under a blanket label. . . .We must avoid blending together in a statistical stew quite diverse problem solving behaviors whose real significance is lost in the averaging process."

Development Steps

Step 1. Describe TS characteristics (in terms of questions).

Global Components. The following set of questions embrace rather global components related to TS (Parker & Dick, 1985):

a. What is the skill level of TS?

b. What is the skill level in using test equipment?

c. What is the skill level in using general-purpose test equipment?

d. Adequacy of specific equipment training; i.e., functional organization and maintenance techniques?

e. What is the basic skill level; i.e., basic subjects such as circuits, schematics, logic gates, etc.?

f. What is the knowledge level of advanced general electronics?

g. What are the basic capacities; i.e., intelligence, verbal, etc.

h. What are the cognitive styles or special aptitudes?

Conceptualization Taxonomy. The following taxonomy translates many of these questions into another conceptualization (Gott, Bennett, & Gillet, 1986; Logan & Eastman, 1986):

a. What are the problem solving skills used in TS; i.e., planning, hypothesis generation, strategies, problem representation?

b. What is the troubleshooters' system understanding or their mental models of a device?

(1) Physical knowledge: physical identity of components.

(2) Functional knowledge: understanding the purpose of a device.

(3) Operational knowledge: understanding the behavior.

(4) System knowledge: understanding the relationship between components and the device as a whole.

c. What is the level of basic procedures or methods such as meter readings, tracing schematics, using test equipment, replacing modules, understanding technical orders and ordnance publications?

Knowledge and Skill Taxonomies (Representations). Numerous taxonomies exist for describing the knowledge and skills of TS (see Denver Research Institute, 1984; Richardson, Keller, Gordon, and Dejong, 1985 for the published proceedings from a Joint Services Workshop on issues of artificial intelligence in maintenance systems including troubleshooting). At a global level, a troubleshooter (a) understands electronic equipment and (b) how to troubleshoot (i.e., solve problems). This is a distinction of a user's device representation and task knowledge representations are explained below:

a. Device representations.

(1) World facts: comprising basic intellectual skills (e.g., language and mathematics) and also knowledge of environmental constraints.

(2) Domain facts: pertain to the field of electronics or related subject domains; for example, principles of electronics (Ohm's law, Kirchoff laws), classes of devices (analogous systems), common test equipment, and similar types of systems.

(3) Device-dependent knowledge.

(a) Physical: identification, descriptive information, and maintenance data such as component failure rates and costs of tests.

(b) Behavioral: specific behavioral information about operating procedures, results of various manipulations, and normal system conditions. Also knowledge of symptom-fault relationships.

(c) Functional knowledge: understanding of how the device works, including design, purpose, and structure of the device.

(d) Unit-specific: concerns the present task; e.g., observable symptoms, complaints, BIT data, or previous maintenance records.

Device knowledge should be organized hierarchically. Often such a model is constructed through analogy with other more observable or familiar systems.

b. Task representation.

(1) Goals: isolate fault, urgent repair, cost effectiveness, etc.

(2) Operators: elementary perceptual, cognitive, or motor activities that change either the technician's mental state or the task environment; i.e., data collection and equipment manipulation activities. For example, take measurements, sensory observations, replace parts, apply signals, trace signal.

(3) Methods: procedures for applying operators, a traditional focus of study.

(a) Symptom-based methods (often the preferred choice of troubleshooters): Possible faults are identified from current symptoms on the basis of associations between various past symptoms and faults. Methods may vary from pattern recognition, trial and error, or classification. These methods are likely to be used in conjunction with more pragmatic concerns such as using least amount of effort, ease of testability, cost, etc.

(b) Specification-based methods (based on more formal analysis of function and structure of a device): Exhaustive search of the device is feasible with relative simple problems and does not require any degree of expertise. Another approach (T-rules) involves mentally simulating a normal system to determine the results of a fault.

(c) Selection rules (heuristics to control the use of various methods according to the task situation's current goals and environmental constraints): Brute force strategies may be adopted when time is limited. S-rules may be applied to routine problems but T-rules are used with complex problems.

This taxonomy suggests that failures in TS may be due to:

(a) Knowledge deficiencies (i.e., mental models).

(b) Recognition failure of critical symptoms and patterns.

(c) Task behaviors (e.g., overconfidence, idiosyncratic, memory limitations).

Training and TS Performance. Morris and Rouse (1985) reviewed several studies that addressed the impact of different training approaches on TS performance. The training approaches incorporated different aspects of job knowledge and skills similar to the distinction afforded by device and task representation as presented by Keller (1985). The four training approaches were:

a. Instruction in the theory upon which the system is based. (These authors seem to be conceptualizing "theoretical knowledge" as the basic principle, laws, functions of some domain; e.g., electronic).

b. Provision of opportunities for TS practice.

c. Guidance in the use of system knowledge. These authors may be using system knowledge as that type of understanding for a particular system; e.g., the Sea Sparrow missile system, while system theory designates basic theoretical domain knowledge.

d. Guidance in the use of algorithms or rules.

Examples for Training Categories. Note that approaches *a* and *b* (above) require that the person develop and use an appropriate strategy while approaches *c* and *d* provide more direct instruction. A brief overview of examples from each category is presented below:

a. Instruction in system theory: Usually this training method is compared to a control group or a group receiving another kind of training.

(1) Shepherd, Marshall, Turner, and Duncan (1977): (a) no story (b) theory (c) heuristics. Heuristics group was better in familiar and unfamiliar situations. Theory group was equivalent to no-story group.

(2) Miller (1975): Theoretical group versus function and action orientation group. Theory group was slower, made more errors, and less successful.

(3) Williams and Whitmore (1959): Theory knowledge was greatest immediately following the training program and lowest on follow-up three years later. TS ability was worst immediately after program and best at follow-up.

(4) Foley (1977): Reviewed seven studies and found that job knowledge tests correlated slightly higher with TS performance than did theory tests.

The focus reported here corresponds to much of the research in other areas of problem solving expertise. That is, explicit training in theories, fundamental, or principles failed to enhance performance of novices. However, the cognitive literature also shows that experts are more proficient in actual problem solving as well as knowledge of domain-specific principles (i.e., device knowledge). The studies listed above did not explicitly investigate the domain-specific knowledge of experts as distinguished between device and task knowledge (cf. Keller, 1985).

b. Opportunity for practice: Morris and Rouse (1985) conclude that practice improves TS performance (speed and accuracy) on both simulations and live equipment. However, they (nor the studies) specifically address why or how practice changes performance (or knowledge). This area needs theoretical development (Schneider, 1985).

c. Guidance in the use of theoretical, context-specific knowledge: These studies have attempted to provide users with general procedures (not specific TS algorithms) to use with their system knowledge. For example, construction of plans and hypotheses, organization of information, hypothesis evaluation, symptom interpretation. The review suggests that these general cognitive procedures are necessary to make effective use of training in system knowledge. Also, these procedures have not been explicitly represented as performance measures in many TS studies. This omission could be addressed by TAE.

d. Guidance in the use of strategies (algorithms, rules, heuristics): General trends suggest that performance can improve with provision of examples, action-related feedback, heuristics, and proceduralization.

Morris and Rouse (1985) conclude that one generic TS ability involves "search for the problem in a systematic manner; in short, to employ some kind of strategy in searching for the source of the difficulty." Strategies can be very direct and overt (and passive) such as specific algorithms or procedures. On the other hand, they can be cognitive procedures (planning, hypothesis generation and evaluation, heuristics), which can vary from general, context-free procedures to more domain-specific ones. Their overall conclusions are: "Either troubleshooters should be explicitly instructed in how to approach problems or they should be forced to use their knowledge of the system explicitly in deciding what to do."

It was proposed that one point of investigation for the TAE Test and Evaluation (T&E) was to assume that proficiency levels embody different qualitative and quantitative levels of knowledge in user's task and device representations. Therefore, measures which will test such representations should be identified and an experimental design which will allow for an adequate check on the validity (criterion and construct) and reliability of the measures should be constructed.

Step 2. Operationally define the constructs and develop measures to test constructs.

Traditionally, the TS research community has focused on product measures such as accuracy and time or on process measures primarily at the level of operators or methods (specifically the symptom-based strategies; cf. Keller, 1985). In other words, much of the research has focused on only the user's task representation. On the other hand, research directed at identifying the user's device representation typically has investigated only general domain facts (e.g., electronic theory) or specific device-dependent knowledge (primarily physical knowledge of descriptive information). Furthermore, this latter emphasis frequently has used paper-and-pencil tests independent of actual job performance situations; in other words, this type of test does not concern itself with evoking a user's selection rules necessary for interfacing device and task representation. With such an approach, the low rate of significant correlations between theory tests and job performance tests is not unexpected. But even with the job performance studies, the focus on user task representations, specifically operator, would not be expected to provide discrimination of TS proficiency since there was no effort to capture the use of methods with respect to a user's device representation.

The TS domain has been the subject of numerous task analyses, primarily of the rational and empirical (behavioral) type. Tasks analyses have taken several forms including observation and interviews of performers, literature reviews of technical documents and manuals, and empirical investigations (including correlational and experimental studies). This body of literature can be surveyed in order to develop multiple indicators and measurements of TS performance. Essentially, TAE can employ these measures to discriminate among levels of TS proficiency. However, two limitations seem to appear if this approach is adopted. First, many of the previous efforts have been primarily concerned with either (a) comparing different TS instructional or training procedures or (b) assessing transfer of training from a variety of instructional materials or procedures to some criterion performance (typically the hands-on or on-the-job test). Individual aptitudes and abilities are often correlated with TS behaviors in either of these two approaches. These approaches have been primarily concerned with issues of criterion validity. As mentioned earlier, this traditional line of research may not go far enough to address issues of construct validity.

Descriptive Measures of Troubleshooting Proficiency. Henneman and Rouse (1984) provide a review of descriptive measures of troubleshooting. These behavioral measures are not equivalent to the specification of TS strategies, heuristics, etc. The implication is that measures can be specified independently of theoretical models of TS; there is some disagreement with this notion; for example, refer to standard philosophy of science discussions (e.g., Kuhn, 1970). There may be a need to consider construct validity of measures which calls for a theoretical analysis (one may argue the same for criterion validity). There does appear to be, however, a strong argument for the behavioral analysis/assessment prior to the cognitive.

The following descriptive measures were used by Henneman and Rouse.

a. Product measures:

   (1) Time: Time to solution (to identify a failed component).

   (2) Cost: Cost of solution (total cost).

   (3) Tico: Product of Time and Cost (speed-accuracy tradeoff).

b. Process measures:

   (1) Number of acceptable actions (ACPN).

   (2) Cost of acceptable actions (ACPC).

   (3) Number of redundant actions (REDN) (did not reduce size of consistent fault set, CFS; i.e., number of plausible faults given the checks obtained thus far; e.g., retesting a component or checking a component whose status could have been inferred from an earlier test) (See Duncan & Shepherd, 1975, for further explanation of CFS).

   (4) Cost of redundant actions (REDC).

   (5) Number of premature replacements (PREN) (could be an acceptable action).

(6) Cost of premature replacements (PREC).

(7) Number of unnecessary actions (UNYN) (when CFS equals one, sufficient information so check is also redundant).

(8) Cost of unnecessary actions (UNYC)

(9) Number of actions that do not use information from a known good component (KNON) (do not use information from components that are known to not be failed).

(10) Cost of above (KNOC).

(11) Average time between actions (AVET) (assess tradeoffs between strategies; e.g., efficiency of tests).

(12) Average time between actions without including "free" actions (AVEG) (e.g., information available without cost).

(13) Average live equipment performance index (EQPP) (assign numerical value to most appropriate actions, neutral, and most inappropriate actions, divide sum by total number of actions).

(14) Average evaluator's rating of live equipment performance (EQPE) (overall subjective index).

Prescriptive Measures Related to Troubleshooting Proficiency. Henneman and Rouse (1984) also specify prescriptive measures which appear related to problem solving skill and TS performance. Consideration of the following (as relevant to area of investigation) is useful for distinguishing expertise and as factors for selection and placement decisions.

a. Cognitive ability:

(1) English: American College Test (ACT) English usage test.

(2) Math: ACT math text.

(3) Social studies: ACT evaluative reasoning, reading, problem solving skill in social sciences.

(4) Natural sciences: ACT natural science test.

(5) Composite of the ACT.

(6) Aviation course grade (basics of aircraft power plants).

(7) Cumulative GPA in aviation courses.

b. Aptitude: Survey of Mechanical Insight (SMI) exam (drawing of a device and multiple-choice questions).

c. Cognitive Style:

(1) Embedded figures test (EFT): distinguishes field dependent and independent; two measures:

(a) Time to solution.

(b) Number of incorrect.

(2) Matching figures test (MFT): distinguishes impulsivity and reflectivity.

(a) Measures of time to first response.

(b) Total number of errors.

Henneman and Rouse computed multiple regression and factor analyses in order to determine the relationship between their descriptive and prescriptive measures. They found that ability scores correlated with one another. SMI did not correlate with any of the ability measures. Also, style measures did not correlate with ability. Thus, style and ability appear to be independent.

Style measures correlated with task (context free) and fault (context free) and fault (context specific) measures (.40). Ability and aptitude measures when considered separately did not correlate with performance measures. However, components of ACT combined with style produced significant regression values (R=.6 to.8).

Three unique dimensions emerged from a factor analysis:

1. Error: COST, TICO, REDN, REDC, UNYN, UNYC, KNON, KNOC.

2. Inefficiency: ACPN, ACPC, -PREN, -PREC (- showing negative correlation to factor).

3 Time: TIME, AVET, AVEG.

Performance Dimensions. It might be useful at this time to compare and contrast these performance dimensions with the measures previously used in the TAE simulations (Conner 1986, 1987); work to determine factors for the TAE RDT&E effort is reported in Conner, Hartley and Mark (1991) with the T&E results.

a. For the time dimension, TAE measures total TS time (no maximum but estimated that the average will be 60 minutes). Also, it seems to be useful and feasible to include other time measures between different actions and tests. Therefore, real time information will be captured for every action of the subjects.

b. For the error dimension, TAE could include some existing measures that quantify the number of invalid checks or other inappropriate actions. These could probably also be assigned cost values.

c. The inefficiency dimension measures actions which are redundant or premature. Relevant TAE measures could include frequency and cost of out-of-bounds test points and invalid checks.

Step 3. Specify the methodology for developing composite score of troubleshooting proficiency.

Attempts to assess job performance proficiency have caused debate over the criterion problem. Any test or task designed to measure proficiency should be evaluated on the basis of its validity; unfortunately, researchers have found that many job proficiency tests do not correlate highly with actual performance or that there may not even be suitable criteria for comparison. A similar criterion problem has been confronted by industrial/organizational psychologists in the job appraisal domain. The controversy has centered on whether an appraisal of an individual's job performance should rely on a composite criterion or upon multiple criteria (Kavanagh, MacKinney, & Wolins,1971; Schmidt & Kaplan, 1971). In either procedure, there is a need for selection, weighting, and combining of job performance elements in order to reach an overall assessment of performance. The appraisal of job performance is critical for two purposes: organizational control and individual development (Hobson, 1981). Organizational control depends on evaluative information in order to make administrative decisions concerning the attainment of goals and purposes. Performance appraisal also serves as an input for issues of training, instruction, and selection of individuals. These limitations and purposes of job appraisal parallel the issues of assessing job proficiency.

The performance appraisal literature suggests that there are serious limitations in the process of obtaining ratings of individual performance (Hobson, 1981). First, supervisors (raters) often are unaware of the number and relative importance of the job dimensions that they use in making overall performance ratings. Second, supervisors often are unable to combine performance information to produce reliable overall ratings. Third, supervisors' subjective (self-report) values of performance information often disagree with the criteria actually used during a performance appraisal. These limitations stem from basic limitations in cognitive processing of complex information as well as various motivational and situational factors. These problems appear in a similar form in the assessment of troubleshooting proficiency. First, researchers have struggled with the criterion problem. Second, any attempt to specify the dimensions of troubleshooting performance and their relative importance may be unreliable for individual assessors and may vary across individuals or groups.

Researchers in the areas of judgment and decision making have developed a methodology known as "policy capturing" that provides an objective, statistical procedure for describing the unique assessment strategies or behaviors of individual raters. Policy capturing involves (a) the presentation to raters of a series of profiles that contains scores on a number of information cues, (b) instructions to raters to review the profile and then assign an overall rating that best summarizes the information, and (c) the use of multiple regression analysis to calculate the extent to which the overall ratings are predictable, given the scores on the cues, and the relative importance of each of the cues in determining the overall rating. The policy equation for an individual, therefore, describes the relationship between each performance dimension and the overall rating.

Judgment Policies/Policy Capturing. For the purposes of TAE, the determination of individual and clusters of judgment policies as applied to the components of TS behavior allow an explicit and objective algorithm for weighting and combining such components into a composite score of TS proficiency. In summary, the identification and comparison of individual policies provide the:

a Extent to which linear and nonliner composite criteria models are employed.

b. Existence of clusters by similarity of policies (and subsequent determination of individual differences among raters).

c. The similarity of statistically derived weights and the decision maker's subjective evaluation of the importance of various criteria (Stumpf & London, 1981).

A study by Zedeck and Kafry (1977) illustrates the policy capturing methodology and related issues of statistical analysis. Zedeck and Kafry used 67 nursing personnel as raters. The group was composed of public health nurses and hospital registered nurses from different supervisory levels and different hospitals. Thus, the subsequent analysis can compare policy equations across different subject groups. Each rater evaluated 40 stimuli (i.e., hypothetical nurses) on the basis of a 7-point scale, reflecting a general global assessment of overall effectiveness. Each hypothetical nurse was described in a paragraph in terms of 9 criterion elements (information cues). For each level, there were 3 possible levels of performance ranging from poor to average to good. Each of the 40 stimuli contained a different combination of levels of the 9 dimensions. Two constraints guided the construction of the stimuli. First, each dimension was approximately normally distributed. Second, the intercorrelations among cues approximated zero. After all of the 40 stimuli were assessed, a rater assigned explicit, subjective weights to each of the 9 cue elements. Each rater also completed a set of scales (e.g., social insight, verbal reasoning, etc.) in order to identify potential correlates of policy clusters.

A multiple regression equation was computed for each rater; R^2 provides an indication of the consistency of the rater's judgment across the 40 stimuli. Relative weights were computed in order to show the proportion of variance contributed by each criterion cue for each rater. Multivariate ANOVAs (MANOVAs) can be computed with the 9 criterion cues as dependent variables. Each relative weight can be compared to the explicit, subjective weights (t-test). Clustering procedures were used to determine clusters of equations that represent groups of raters. These clusters can be analyzed by comparison to background variables (e.g., aptitude and personality measures). A MANOVA can be computed to determine differences between the clusters in terms of the criterion cues. Discriminant analyses can be computed with individual differences measures as predictors and cluster membership as a criterion measure.

Policy capturing studies demonstrate that there may be considerable individual differences among raters. This methodology provides a basis for identifying such differences and a starting point for resolution of such differences. Performance assessment would benefit from the identification of individual differences since different policies can be used to compare the relative advantages/disadvantages of different strategies of performance assessment. Furthermore, studies show that raters' explicit, subjective weighting strategies may deviate considerably from the implicit, objective weightings actually used in performance assessment. Finally, the purpose of

the ratings can be varied in order to determine how policies are likely to change as a function of assessment context (e.g., promotion, selection, training, evaluation).

METHODOLOGY FOR TAE TEST AND EVALUATION

Determination of Troubleshooting Proficiency Composite Score

Assumptions

1. The initial focus of the TAE T&E was to develop a set of measures that discriminate between levels of TS proficiency and demonstrate validity and reliability.

2. Initially, the assessment of TS proficiency focused on behavioral discriminations; i.e., the assessment consisted of typical product and process measures (Henneman & Rouse, 1984) obtained from behavior on simulated TS tasks.

3. The main assessment technique to be used was the microcomputer-delivered scenarios developed for TAE. Therefore, performance measures were limited to those variables that a microcomputer can collect and store. These measures are representative of the troubleshooter's task knowledge (Keller, 1985) specifically, "operators" and "methods."

4. The search for an "ultimate" criterion for the TAE test may be constrained by several problems in typical job proficiency and performance tests (e.g., Pickering & Bearden, 1984); therefore, discrimination of TS proficiency will depend upon the development of a composite score to be formed from the TAE performance measures.

5. The composite score should be derived from an objective and reliable combination of the ranked and weighed performance factors such that this composite score will be applicable across different TS scenarios and contexts. The algorithm for computing the composite score should be developed from information obtained from a representative sample of individuals with expertise in the domain of electronics maintenance and troubleshooting.

6. Furthermore, the algorithm for computing the composite score should be modifiable in order to represent different weighting of the performance measures. These differences will be sensitive to the different perspective from individual experts in the domain.

7. One method for obtaining the algorithm for TS composite scores is to modify the policy-capturing procedures discussed previously. The policy-capturing methodology can be used in the TAE evaluation to achieve the general goal of developing a scoring algorithm which weigh and combine the different TS components into a composite proficiency score. Thereafter, proficiency scores can be used to assess an individual's performance across TS scenarios and also to compare the proficiency levels of different individuals. Additionally, the composite score will provide an objective and constant assessment procedure in order to reveal the development of an individual's TS proficiency over time and to compare the effect of different training and instructional interventions on TS proficiency.

The procedures for developing a TS proficiency composite score are outlined below.

Procedures

1. Generate a list of potential cue components relevant to measures of troubleshooting proficiency. Given the assumptions (above), the cue components will be limited to the behavioral measures that can be captured by the microcomputer-delivered system. These cue components can consist of: found solutions, proof points, checks, invalid checks, out-of-bounds checks, illogical approach, incorrect solutions, total time, time to first proof point, time per check, etc.

2. Develop the presentation context of the cue profiles which could consist of two forms:

a. Give the expert judge (rater) the symptoms from a scenario selected from earlier TAE research and have the judge troubleshoot the scenario. Slight variation here could be to give symptoms and fault so that the judge can quickly search the problem space (i.e., the path between symptoms and fault). This format represents a scenario-context approach.

b. The expert judge receives cue profiles without reference to any particular scenario. This format represents a context-free approach.

In either presentation context, the judge will see various profiles containing combinations of cue components. Typically, policy-capturing studies have included from 3 to 15 cue components. Obviously, how information processing limitations, motivation, fatigue, etc. will determine the upper limit of number of cues given in a profile needs to be discussed.

3. Once cue components have been selected, it is necessary to determine whether cue scores should be presented as:

a. Raw data.

b. Scaled scores.

This decision will also be related to the choice of presentation context. Raw scores would seem more realistic if raters receive an actual set of symptoms for a scenario. Some components may require information such as two of four proof points tested. Scaled values (e.g., ranging from unacceptable to acceptable) may be more appropriate for a context-free presentation such as percentiles or Z-scores.

4. There may be two profile formats to choose from:

a. One profile format could consist of a sequence of steps as summarized in TAE (e.g., check, check, proof point, check, incorrect solution, proof point, solution). In this format, the rater has both summative information (i.e., number of checks, etc.) but also the sequence of steps. Certain sequencing patterns in this case should be considered as cue components; i.e., reflective of different strategies, their efficiency, etc.

b. Another profile format could be similar to the scoring presentation used in TAE (without the associated point values).

5. The number of profiles to be constructed and presented to each judge must be determined. Typically, the number should increase with an increase in the number of cues. The construction of profiles (i.e., the cue component and value matrix) can vary along several dimensions:

a. Actual subject performance in a particular scenario could be used as the basis of the profiles; i.e., each (of a sample) subject's performance would comprise a separate profile (the format being either the sequence of steps or the scoring summary discussed in Step 3). In this case, the cues would be intercorrelated; there are some methodological arguments whether cue intercorrelations should approximate zero. Statistically, zero intercorrelations may increase multiple-R but the composition of profiles may be unrealistic of the actual situation of cue relationships.

b. Scenarios could be constructed by algorithmic procedures to ensure that cue intercorrelations correlate near zero.

6. The overall rating scale and behavioral anchors need to be chosen. Typically, this scale has anchors to communicate some overall assessment. The purpose of assessment may be manipulated in order to determine its effect on policy strategy.

7. Rater groups need to be determined and selected to compare different groups in terms of experience, context, training, etc.

Policy Capturing Example

Once the TAE has performance results, the following example illustrates the procedural choices of context-free, combination of scaled scores and raw scores; scoring summary profile format; uncorrelated cue value matrix; and different groups of judges:

1. Procedures

a. Each judge will be tested separately. The judges are told that they will see 50 different scoring profiles obtained from actual TS performances of "C" school qualified subjects. (One option is to describe the type and nature of the TS task; i.e., the particular hardware and the TS context such as whether the ship is at sea, in dock, etc.). They are to examine each profile in order to reach an assessment of the overall troubleshooting proficiency on that given scenario. A profile may assume the following form:

Scaled scores range from 1 = very unfavorable to 9 = very favorable

Found solution: Yes (or No)

Proof Points: 2 of 4

Checks: 5

Invalid Checks: 4

Out of Bounds: 2

Illogical Approach: 4

Incorrect Solutions: 1

Time to First Proof Point: 15 minutes

Average Time per Check: 2 minutes

Total Time: 20 minutes

Overall Assessment:

Very Unacceptable ___ ___ ___ ___ ___ ___ ___ Very Acceptable
1 7

The judges are told to read the scoring profile and to place the profile into one of the seven possible rating levels so that number of profiles in a group approximate the following normal distribution:

1- ?

2- ?

3- ?

4- ?

5- ?

6- ?

7- ?

b. After all profiles have been rated, the judges will be told to consider the number of criterion elements and to divide 100 points among the elements to reflect the relative importance they hold in proficient troubleshooting. This information will be analyzed to identify each judge's explicit, subjective weights.

c Biographical data of the judges will then be collected (e.g., experience, cognitive style tests, etc.).

2. The data are analyzed and used in the following ways:

a. Individual multiple correlation coefficients will indicate the consistency of each judge's policy for rating troubleshooting proficiency across several different performances. If the judge's ratings are not consistent enough, the composite algorithm may not meet the reliability demands.

b. The weights of the criterion elements will reveal the implicit, but objective importance of the components used to reach the assessment criterion. These weights provide an objective basis for the derivation of a scoring algorithm for computing subsequent troubleshooting composite scores. Also, the equation may reveal that certain criterion elements have very little relative importance and hence their measurement may not be necessary.

c. The objective weights of any judge can be compared to the same judge's subjective, but explicit, weights of criterion elements. Differences should help reveal inconsistencies in the judge's performance ratings. In addition, the scoring algorithm from this method may be compared to algorithms derived from other methods.

d. Analyses of the regression equations across judges can reveal clusters of ratings that define separate groups of judges. This information may reveal that any overall proficiency score may differ as the weightings change for different audiences; e.g., instructors versus technicians.

e. By capturing judges' policies across different troubleshooting scenarios, it is possible to determine if a troubleshooting proficiency composite score can be applied across scenarios or if there is a basis for scenario specificity.

The following studies contain examples of methodological procedures and data analyses: Arima & Neil, 1978; Borman & Dunnette, 1974; Christal, 1968; Hobson, 1981; Slovic & Lichenstein, 1971; Stumpf & London, 1981; Zedeck & Kafry, 1977.

Assessment of Empirical Validity of TAE Composite Score

Assumptions

1. Although the TS composite score should reveal reliable differences in levels of proficiency, there also should be assessment of its empirical validity.

2. Empirical validity may be assessed by comparing TS composite scores with other, traditional measures of TS performs although the limitations of such measures should be remembered (e.g., Arima & Neil, 1978; Pickering & Bearden, 1984; Vineberg & Joyner, 1982).

Procedures

Three alternative procedures are possible.

1. Appropriate evaluators (e.g., school instructors) may be given a student's TAE scoring summary for each scenario completed. The evaluator will rate the student's overall TS proficiency on each scenario or all scenarios. These evaluator ratings will be correlated with the student's composite score(s) in order to assess the degree of relationship between the two types of assessments.

2. A panel of judges can determine the "optimum path" (strategy) for proficiently solving the scenario problem. The sequence of steps taken by a student on the microcomputer-delivered scenario can be compared to the optimum and any deviation can be evaluated. The evaluative

interpretation of the deviation can be compared to the TS composite score in order to determine their relationship.

3. Student TS composite score on selected scenarios delivered by the microcomputer can be compared to student performance on the same scenario hardware. Measures to be obtained from the actual TS performance can include:

a. Instructor's rating.

b. Comparison of the component measures (e.g., number and type of checks, out of bounds) to those of the student on the microcomputer delivered scenario.

Determination of Behavioral Factors Underlying Levels of TS Proficiency

Assumptions

1. "C" school students receive several TS scenarios by the microcomputer-delivered TAE technique. Composite TS proficiency scores will be computed by the algorithm developed from the policy-capturing study. Proficiency scores will be available for each student on each scenario.

2. Analyses and subsequent normative comparisons of TS composite scores should reveal individual differences in troubleshooting proficiency.

3. However, since the composite score represents a weighted combination of separate (although possibly correlated) performance measures, additional measures and analyses should be undertaken in order to reveal individual differences in the performance measures as well as other related factors such as experience, ability, aptitude, and knowledge. These analyses will serve as the initial attempts to specify the behavioral, cognitive, and perceptual components of TS proficiency.

4. Appendix A lists the measures which have been used in TS research. Henneman and Rouse (1984) found three factors underlying descriptive measures of TS behavior:

a. Error.

b. Inefficiency.

c. Time.

These factors may also distinguish between levels of TS proficiency found in the TAE T&E.

Procedures

1. Factor analysis of the separate TS measures captured by the microcomputer-delivered TAE technique can reveal the intercorrelation of measures and any underlying factor structure.

2. Factor scores can be compared across individuals with different levels of overall proficiency in order to determine their relative contribution.

3. Several individual prescriptive measures (cf. Henneman & Rouse, 1984) can be obtained from the "C" school students who are assessed by the TAE microcomputer technique. These measures can also be factor analyzed to determine their factor structure and also serve as predictor variables to be used in multiple regression analyses with TS composite scores or performance factors as criterion variables.

Determination of Complex Cognitive Skills Underlying TS Proficiency

Assumptions

1. Differences in TS proficiency (as revealed by behavioral measures) are representative of a multifaceted competence (behavioral, cognitive, and perceptual aspects).

2. Explanation of TS proficiency as well as instructional, training, and simulation endeavors to improve TS proficiency will benefit from the explication of underlying competencies as revealed by a cognitive analysis (Glaser, 1976).

3. Any behavioral performance of TS will depend upon the knowledge and skills evoked during such performance. This underlying competence can be distinguished as two types of knowledge representations: device representation and task representation (cf. Gott et al., 1986; Keller, 1985; Richardson et al., 1985).

4. The TS composite score is representative primarily of troubleshooter's task representation, specifically operators and methods. Hence, further data collection procedures and analyses should be planned to explicate *symptom-based versus specification*-based methods (cf. Rasmussen, 1983, and his distinction of S-rules and T-rules) and device-dependent knowledge (i.e., physical, behavioral, functional, and unit-specific).

5. Often, expert informants are very reluctant to accept the contention that their expertise lies in such "abstract, elusive, phenomenological, mental entities" (i.e., mental models, cf. Gott et al., 1986). Much of their expertise is more directly available to them in the form of empirical associations. That is, they have largely restructured their conceptual knowledge (cf. Gott et al., 1986) into efficient pattern recognition skills which may only be verbalized in crude rule-like associations (e.g., if X, then Y). Of course, these rules provide the expert with a vast repertoire of domain-specific troubleshooting knowledge. The problem is that experts often claim that such knowledge has been gathered from their available experience and cannot be taught (because of numerous context-specific factors).

6. Another issue is that experts have automatized many basic behavioral and cognitive procedures into larger units or actions. Since this knowledge is compiled, TS performance achieves a great deal of efficiency in addition to proficiency. But compiled knowledge is not always readily available (or may be subject to various sources of distortion when verbalized). So the researcher must devise tasks that will "unpack" or "de-compile" such expert knowledge. Cognitive tasks analyses have adopted the use of tasks which are not directly measuring as performance. However, any attempt to have experts respond to "non- task" (job) problems may further alienate the expert who will claim that these tasks do not tap their expertise.

Procedures

1. Small groups of subjects are chosen to provide representative scores obtained from the microcomputer-delivered TAE technique. Additional groups may be selected to represent other levels of expertise or experience on the basis of rank, level, experience, or peer ratings.

2 Tasks along the lines of those used by Glaser (1985) and Gott et al., (1986) were developed and administered to subjects. These tasks require the collection of "thinking-aloud" verbal protocols which are tape-recorded and then converted into typewritten transcripts. A verbal-protocol analysis can be undertaken to explicate differences in the conceptual understanding and functional knowledge which correspond to the identified differences in TS behavior as revealed by the TAE composite score.

3. Specification of mental models (i.e., task and device representations) that correspond to different competence levels of TS will serve as input to the development and implementation of intelligent tutoring systems (e.g., Gott et al., 1986), training simulators (Montague, 1982), intelligent maintenance aids (e.g., Keller, 1985) and automated systems for hardware maintenance (e.g., Denver Research Institute, 1984; Richardson et al., 1985).

RECOMMENDATIONS FOR TAE DEVELOPMENT

A number of recommendations for TAE development efforts are presented. Appendix B presents the test and evaluation plan derived from t these recommendations.

1. Determine, based on expert knowledge of instructors, experienced shipboard technicians and other personnel (perhaps engineers), the possible failures of the systems under investigation:

a. Determine both the common and less frequent problems which prevent the equipment from functioning properly; use expert opinion as well as any records maintained aboard ships or any other maintenance information system.

b. Define the problems in terms of symptoms as well as causes.

2. Determine, through expert opinion and engineering data, the optimum (most efficient, theoretically most correct) series of steps that can be used to diagnose and solve the problem given the nature of the equipment, testing procedures, and aids and training.

3. Insert, into the equipment, a series of faults that represents the various equipment problems that do or may occur. Include difficult or rare problems to obtain a representative sample.

a. Carefully observe and record the behaviors of a sample of expert, experienced maintenance personnel (including instructors) troubleshooting the various inserted faults.

b. Score the performance of the experts using the various dimensions of performance found in the literature and/or determined by a jury of experts and subject matter experts including:

(1) Accuracy of diagnosis.

(2) Time to diagnosis.

(3) Various types of errors and inefficiencies.

(4) Overall sequence of steps.

(5) Quality of the first step.

(6) Responses to incorrect hypotheses or tests.

(7) Comparison to the theoretically optimal or ideal solution.

(8) Costs in terms of time, errors, parts replaced, tests made.

(9) Use of information in terms of its optimum - what is the best step knowing what will most quickly and accurately lead to the solution.

It may be useful to have a subsample of expert troubleshooters verbalize what they are doing, or possibly video tape their activities and actions, during the initial tests of the scenarios to gain some insight into the mental (cognitive) processes that are taking place.

c. Develop composite and part scores for each fault inserted in the equipment. The statistical relationship between the part scores (e.g., errors, time, overall strategy, degree of efficiency, number of checks, costs) and the overall composite score (developed by the experts) should be established. An alternative to developing a composite score, taking into account all measured behaviors, would be to deal with outcome measures. Outcome measures would consist of the gross products of the problem solving process such as number of problems solved correctly, average time needed to solve the problems and deviation from standards of optimum (or best) performance based on performance of a sample of expert technicians. Process measures, both qualitative and quantitative, can be assessed as individual variables. The interrelations among the process measures and the relationships between the process and outcome measures would provide important information. It is possible that the results of these analyses will produce a limited number of process measures such as a series of factors. Scores on these factors should be related to outcome measures. This alternative would not be oriented to one weighted composite measure of performance but would result in a limited number of outcomes and process measures. A composite, although useful for analytic, selection, and prediction purposes, may limit understanding and insight that may be revealed through separate measures of process and outcomes.

d. Determine the consistency of performance.

(1) Consistency of the individual over the various episodes measured by the composite as well as by the various measures suggested above (Step 3b).

(2) Consistency of the sample of experts for each episode or logical category of episodes.

(3) The degree of both qualitative and quantitative consistency within individuals (over episodes) and by episodes (over individuals) will determine standards for performance. If the

degree of consistency in terms of overall performance or process performance appears low, one must re-think the scoring procedures, the episodes inserted in the equipment or whether or not the troubleshooting behavior is consistent.

4. If standards of performance can be developed from expert opinion and expert performance, the TAE microcomputer-based test, will enable an assessment of the discrepancies between technicians, performance, and the best solution that can be achieved by experts. This information will be useful as feedback to the schools as well as for developing refresher or remedial training.

5. The microcomputer scenarios should follow the types of episodes found in the actual equipment including episodes that are not directly dealt with in training, but that can be solved by novices who have learned a strategy based on equipment knowledge (symptoms and causes) and how to obtain and use information through tests and checks.

6. Estimate reliability of the computer-based test consisting of a standard set of scenarios through various statistical methodologies using gross outcome measures (number solved correctly, number solved correctly within time limits, costs, etc.) as well as the individual process measures (number of errors, overall sequence of actions, average degree of reduction of the consistent fault set, amount of information obtained for each action, etc.). The test should contain enough problems to make reliability estimates of gross performance (number solved correctly within an allocated amount of time).

7. Content validity can be obtained by selecting the types of episodes that are representative of problems found aboard ship as well as varying the degree of difficulty. The degree of difficulty can be estimated by consistent differences among the various episodes with respect to how many individuals solve the episode within the time constraints. Concurrent validity can be estimated by relating scores on the test to school performance, ratings by supervisors aboard ship, and similar measures.

8. Criterion or predictive validity is difficult to define. Scores (process and outcomes) on the TAE computer-based test will be validated against future criteria of performance including:

a. Performance aboard ship on actual troubleshooting tasks. This criterion may be difficult to establish. Troubleshooting cannot be determined in advance and the conditions under which breakdowns occur are not standard nor predictable. Thus, the criterion is uncertain and performance may be limited in terms of the range of troubleshooting activities as well as frequency. Conditions may vary. Appropriate and similar performance opportunities aboard ship may vary greatly among technicians.

b. Supervisory ratings. The use of supervisory ratings as a criterion has many and varied weaknesses.

c. Testing technicians aboard ship with the computer-based test using different scenarios than used in the original test. If this is feasible for enough tested technicians aboard a variety of ships, the predictive validity can be estimated.

d. Testing on the actual equipment aboard ship with known and representative faults inserted and measures of fault diagnosis, parts replacement (if needed), and time to solution carefully measured under standard conditions. Scores on both process and outcome performance measures from the TAE test can be related to performance on the actual equipment with known faults. This form of testing may be impossible due to safety requirements and the need to maintain the equipment in a state of operational readiness. Among these measures are supervisory ratings at various time periods after "C" School graduation, promotions, commendations, and other indicators of technical skill performance.

e. It may be necessary to obtain criterion (or predictive) validation by means of several measures, none of which are ideal and each of which may, in some way, assess part of the complex set of abilities involved in troubleshooting. These measures may include supervisor ratings over time, promotions, commendations and other organizational indicators of an individual's technical skills.

REFERENCES

Arima, J., & Neil, D. (1978). *Skill deterioration and its management* (NPS 55-78-7). Monterey, CA: Navy Postgraduate School.

Ilhisi, E. C. (1977). Performance measurement technology: Issues and answers. In *Proceedings of symposium on productivity enhancement: Personnel performance assessment in systems* (pp. 343-360). San Diego: Navy Personnel Research and Development Center.

Borman, W. C., & Dunnette, M. D., (1974). *Selection of components to comprise a Naval perspective status index (NPSI) and a strategy for investigating their relative importance.*Washington DC: Office of Naval Research. (AD-776 285))

Bryan, G. L. (1954). *The AUTOMASTS: An automatically recording test of electronics troubleshooting* (Electronics Personnel Research Group Report No. 11). Los Angeles: University of Southern California, Department of Psychology.

Christal, T. (1968). Selecting a harem and other applications of the policy capturing model. *The Journal Of Experimental Education, 36*(4), 35-41.

Conner, H. B. (1986, October). Troubleshooting proficiency evaluation project (TPEP). In *Proceedings of Military Testing Association Conference*, Mystic, Connecticut.

Conner, H. B. (1987, April). Troubleshooting Proficiency Evaluation Project (TPEP) for the NATO Seasparrow Surface Missile System (NSSMS): A feasibility study. In *Proceedings of First International Manpower and Training Conference of the National Security Industrial Association*, Luxembourg.

Conner H. B., Hartley, S., & Mark, L. J. (1991). *Troubleshooting Assessment and Enhancement (TAE) Program: Test and evaluation* (NPRDC-TN-91-13). San Diego: Navy Personnel Research and Development Center.

Conner, H., Poirier, C., Ulrich, R., & Bridges, T. (1991). *Troubleshooting Assessment and Enhancement (TAE) Program: Computer delivery system design, development, and program administration* (NPRDC-TN-91-12). San Diego: Navy Personnel Research and Development Center.

Demaree, R. G., Crowder, N. A., & Morrison, E. J. (1955). *Proficiency of Q-24 radar mechanics: Summary of findings* (AFPRTC T TM-55-6). Lowry Air Force Base, CO: Air Force Personnel and Training Center.

Denver Research Institute (1984). *Artificial intelligence in maintenance: Proceedings of the Joint Services Workshop* (AFHRL-TR-84-25). Lowry Air Force Base, CO: Air Force Human Resource Laboratory.

Duncan, K. D., & Shepherd, A. (1975). A simulator and training technique for diagnosing plant failures from control panels. *Ergonomics, 1975, 18*, 627-641.

Foley, J. P., Jr. (1974). *Evaluating maintenance performance: An analysis* (Technical Report 74-57(1)). Wright Patterson Air Force Base, OH: Air Force Human Resource Laboratory.

Foley, J.P., Jr. (1975). *Criterion-referenced measures of technical proficiency in maintenance activities* (Technical Report AFHRL-TR-75-61). Wright Patterson Air Force Base, OH: Air Force Human Resource Laboratory.

Foley, J. P., Jr. (1977) *Performance measurement of maintenance* (Technical Report AFHRL-TR-77-76). Wright Patterson Air Force Base, OH. Air Force Human Resources Laboratory.

Glaser, R. (1976). Components of a psychology of instruction: Toward a science of design. *Review of Educational Research, 46*(1), 1-24.

Glaser, R. (1985). *Cognitive task analysis to enhance technical skill training and assessment.* Pittsburgh, PA: University of Pittsburgh, Learning and Research Development Center.

Gott, S., Bennett, & Gillet (1986). Models of technical competence for intelligent tutoring systems. *Journal of Computer-Based Instruction, 13* , 43-46.

Harris, J. H., Campbell, R. C., Osborn, W. C., & Boldovici, J. A. (1975). *Development of a model job performance test for combat occupational specialty. Volume II: Instructions and procedures for conducting a functionally integrated performance test* (Final Report FR-CD (L)-75-6-V1). Alexandria, VA: Human Resources Research Organization.

Henneman, R. L., & Rouse, W. B. (1984). Measures of human problem solving performance in fault diagnosis tasks. *IEEE transactions on Systems, Man, and Cybernetics, SMC-14,* 99-112.

Hobson, C. (1981). Clarifying performance appraisal criteria. *Journal of Organizational Behavior and Performance, 28,* 164-188.

Kavanagh, M. J., MacKinnery, A.C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75,* 34-39.

Keller, R. (1985). *Human troubleshooting in electronics: Implications for intelligent maintenance aids* (AFHRL-TR-85-34). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Kieras, D. E., & Polson, P. G. (1982). *An approach to the formal analysis of user complexity* (Working paper No. 2). Tucson AZ: University of Arizona.

Kuhn, T. S. (1970). *The structure of scientific revolution* [2nd Ed. Chicago, Univ. of Chicago Pr., 1970] (Intl. Encl. of Unified Sci. Foundation of the Unity of Sci, V.2, No.2).

Logan, D., & Eastman, R. (1986). *Mental models of electronic troubleshooting.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Mackie, R. R., McCauley, M. E., & O'Hanlon, J. F. (1978). *New criteria for selection and evaluation of sonar technicians. Phase III: Trial administration of experimental prediction tests* (NPRDC TN-78-13). San Diego: Navy Personnel Research and Development Center.

Mallory, W. J., & Elliott, T. K. (1978) *Measuring troubleshooting skills using hardware-free simulation* (AFHRL-TR-78-47). Lowery Air Force Base, CO: Air Force Human Resources Laboratory.

Miller, E. E. (1975). *Instructional strategies using low-cost simulation for electronic maintenance* (Tech. Rept. HumRRO-FR-WD(TX)-75-20). Alexandria, VA: Human Resources Research Organization.

Montague, W. (1982). *Is simulation fidelity the question* (NPRDC TN 82-13). San Diego: Navy Personnel Research and Development Center.

Morris, N. M., & Rouse, W. B. (1985). Review and evaluation of empirical research in troubleshooting. *Human Factors, 27,* 503-530.

Nauta, F. (1985). *Alleviating fleet maintenance problems through maintenance training and aiding research* (NAVTRAEQUIPCEN MDA903-81-C-0166-1). Orlando, FL: Naval Training Equipment Center.

Nunnally, J. C., & Durham, R. L. (1975). Validity, reliability and special problems of measurement in evaluation research. In E. L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1, pp. 289-352). Beverly Hills, CA: Sage Publications.[1]

Parker, E., & Dick, R. (1985). *Factors responsible for Navy electronic equipment downtime: Recommendations for corrective research and development* (HFOSL TN-71-85-08). San Diego, CA: Navy Personnel Research and Development Center.

Pickering, E. J., & Anderson, A. V. (1976). *A performance-oriented electronics technician training program: I. Course development and implementation* (Technical Bulletin 67-2). San Diego, CA: Navy Personnel and Training Research Laboratory.

Pickering, E., & Bearden, R. (1984). *Job-performance testing research at the Navy Personnel Research and Development Center--1953 through 1981* (NPRDC TR-84-36). San Diego: Navy Personnel Research and Development Center. (AD-B081 904L)

Rasmussen, J. (1983). Skills, rules, and knowledge: Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics,* 13(3), 257-266.

Richardson, J.J., Keller, R. A., Polson, P. G., & Dejong, K. A. (1985). *Artificial intelligence in maintenance synthesis of technical issues* (AFHRL TR 85-7). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

---

[1]Cited in Appendix B.

Rose, A. M., Fingerman, P. W., Wheaton, G. R., Eisner, E., & Kramer, G. (1974). *Methods of predicting job-ability requirements: II. Ability requirements as a function of changes in the characteristics of an electronic fault-finding task* (Tech. Rept. R74-6). Washington, DC: American Institute for Research.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24*, 419-434.

Schneider, W. (1985). Training high-performance skills: Fallacies and guide lines. *Human Factors, 23*(3).

Shepherd, A., Marshall, E. C., Turner, A., & Duncan, K. D. (1977). Diagnosis of plant failures from a control panel: A comparison of three training methods. *Ergonomics, 20*, 347-361.

Shriver, E. L., & Foley, J. P., Jr. (1974, November). *Evaluating maintenance performance: The development of graphic symbolic substitutes for criterion referenced job task performance tests for electronic maintenance* (Tech. Report AFHRL-TR-74-57). Wright Patterson Air Force Base, OH: Air Force Human Resources Laboratory.

Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance, 6*, 649-744.

Stumpf, S., & London, M. (1981). Capturing rater policies in evaluating candidates for promotion. *Academy of Management Journal, 24*, 752-766.

Vineberg, R. A. (1968, June). A performance test for the AAFCS M33 radar mechanic and observations on troubleshooting behavior (Paper for Symposium on Electronics Maintenance, Office of Assistant Secretary of Defense Research, and Development, Advisory Panel on Personnel and Training Research). In *Collected papers prepared under work unit RADAR* (HumRRO Prof. Paper 20-68). Alexandria, VA: George Washington University, Human Resources Research Office.[1]

Vineberg, R. A., & Joyner, J. (1982). *Prediction of job performance: Review of military studies* (NPRDC TR-82-37). San Diego: Navy Personnel Research and Development Center. (AD-A113 208)

Vineberg, R., & Taylor, E. (1972). *Performance in four Army jobs by men at different aptitude levels: 4. Relationship between performance criteria* (TR 72-33). Alexandria, VA: Human Resources Research Organization.

Wetzel, S. K., Konoske, P. J., & Montague W. E.(1983). *Effects of display format on sonar operator performance* (NPRDC SR-83-37). San Diego: Navy Personnel Research and Development Center. (AD-A129 496)

[1]Cited in Appendix B.

Williams, W. L., Jr., & Whitmore, P. G., Jr. (1959). *The development and use of a performance test as a basis for comparing technicians with and without field experience: The Nike Ajax maintenance technician* (TR 52). Washington, DC: George Washington University, Human Resources Research Office.

Zedeck, S., & Kafry, D. (1977). Capturing rater policies for processing evaluation data. *Organizational Behavior and Human Performance, 18*, 269-294.

BIBLIOGRAPHY

Abrams, A., & Pickering, E. J. (1963). *Experimental training of sonarmen in the use of electronic test equipment. V. Performance test results on a diagnostic troubleshooting test* (Technical Bulletin 63-15). Washington, DC: Bureau of Naval Personnel.

Baldwin, R. D. (1978). *Training the electronics maintenance technician* (HumRRO Professional Paper 7-78). Alexandria, VA: Human Resources Research Organization.

Bond, N. A., Jr., & Rigney, J. W. (1966). Bayesian aspects of trouble shooting behavior. *Human Factors, 8*, 377-383.

Bottenberg, R. A., & Christal, R. E. (1966). Grouping criteria - A method which retains maximum predictive accuracy. *The Journal of Experimental Education, 36*, 28-34.

Brigham, R., & Laios, L. (1975). Operator performance in the control of a laboratory process plant. *Ergonomics, 18*, 53-66.

Brooke, J. B., Cooke, J. F., & Duncan, K. D. (1983). Effects of computer aiding and pretraining on fault location. *Ergonomics, 26*, 669-686.

Brooke, J. B., & Duncan, K. D. (1981). Effects of system display format on performance in a fault location task. *Ergonomics, 24*, 175-189.

Brooke, J. B., Duncan, K. D., & Cooper, C. (1980). Interactive instruction in solving fault-finding problems--an experimental study. *International Journal of Man-Machine Studies, 12*, 217-227.

Brooke, J. B., Duncan, K. D., & Marshall, E. C. (1978). Interactive instruction in solving fault-finding problems. *International Journal of Man-Machine Studies, 10*, 603-611.

Cicchinelli, L. F., Keller, R. A., & Harmon, K. R. (1984). *Training capabilities test of Electronics Equipment Maintenance Trainer (EEMT): Findings and conclusions* (Draft report). Denver CO: Denver Research Institute, University of Denver.

Cooke, J. F., & Duncan, K. D. (1983). Training on fault location. *Ergonomics, 26*.

Cornell, F. G., Damrin, D. E., Saupe, J. L., & Crowder, N. A. (1954). *Proficiency of Q-24 radar mechanics: III. The tab test--a group test of trouble-shooting proficiency* (AFPTRC-TR-54-52). Lackland Air Force Base, TX: Air Force Personnel and Training Center Research Bulletin.

Dale, H.C.A. (1958). Fault-finding in electronic equipment. *Ergonomics, 1*, 356-385.

Duncan, K. D. (1971). Long-term retention and transfer of an industrial search skill. *British Journal of Psychology, 62*, 439-448.

Duncan, K. D. (1981). *Training for fault diagnosis in industrial process plants.* In J. Rasmussen & W. B. Rouse (Eds.), *Human detection and diagnosis of system failures*. New York: Plenum.

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical problem solving*. Cambridge, MA: Harvard University Press.

Federico, P-A. (1983). Changes in the cognitive components of achievement as students proceed through computer-managed instruction. *Journal of Computer-Based Instruction, 9,* 156-168.

Federico, P-A. (1982). Individual differences in cognitive characteristics and computer-managed mastery learning. *Journal of Computer-Based Instruction, 9,* 10-18.

Federico, P A., & Landis, D. B. (1979). *Predicting student performance in a computer-managed course using measures of cognitive styles, abilities, and aptitudes* (NPRDC TR-79-30). San Diego: Navy Personnel Research and Development Center. (AD-A074 880)

Federico, P-A., & Landis, D. B. (1980). *Relationships among selected measures of cognitive styles, abilities, and aptitudes* (NPRDC TR-80-23). San Diego: Navy Personnel Research and Development Center. (AD-A090 729)

Finch, C. R. (1971). *Troubleshooting instruction in vocational technical education via dynamic simulation.* Pennsylvania Dep. Education Project, Rep. 10064, 19-1024.

Finch, C. R. (1972). The effectiveness of selected self-instructional approaches in teaching diagnostic problem solving. *Journal of Educational Research, 43,* 219-222.

Ginsburg, A. (1985). Comparison of intratraining evaluation with test of clinical ability in medical students. *Journal of Medical Education, 60,* 29-36.

Glaser, R., & Phillips, J. C. (1954, August). *An analysis of proficiency for guided missile personnel: III. Patterns of troubleshooting behavior* (Tech. Bulletin 55-16). Washington, DC: American Institute for Research.

Glass, A. A. (1967). Problem-solving techniques and troubleshooting simulators in training electronic repairmen (Doctoral dissertation, Columbia University). *Dissertation Abstracts International.* (University Microfilms No. 67-12, 252)

Gruppen, L., & Wolf, F. (1985). Expertise, efficiency and the construct validity of patient management problems. *Journal of Medical Education, 60,* 878-880.

Highland, R. W., Newman, S. E., & Waller, H. S. (1956). A descriptive study of electronic trouble shooting. In *Air Force Human Engineering, Personnel, and Training Research* (Tech. Rept. 56-8). Baltimore, MD: Air Force Research and Development Center.

Hunt, R. M., Henneman, R. L., & Rouse, W. B. (1981). Characterizing the development of human expertise in simulated fault diagnosis tasks. In *Proceedings of the International Conference on Cybernetics and Society* (pp. 369-374). Atlanta, GA: IEEE Systems, Man, Cybernetics Society.

Hunt, R. M., & Rouse, W. B. (1981). Problem-solving skills of maintenance trainees in diagnosing faults in simulated powerplants. *Human Factors, 23,* 317-328.

Johnson, W. B., & Rouse, W. B., (1982) Training maintenance technicians for troubleshooting: Two experiments with computer simulations. *Human Factors, 24*(3), 271-276.

Kieras, D. E., & Polson, P. G. (1982). *An outline of a theory of the user, complexity of devices and systems* (Working paper No. 1). Tucson, AZ: University of Arizona.

Laabs, G. C., Main, R. E., Abrams, A. J., & Steinemann, J. H. (1975). *A personnel readiness training program: Initial project development* (NPRDC SR-75-8). San Diego: Navy Personnel Research and Development Center.

Landeweerd, J. (1979). Internal representation of a process, fault diagnosis and fault correction. *Ergonomics, 22,* 1343-1351.

Magone, M., & Yengo, L. (1986, April). *A conceptual framework for technical skills training.* Paper presented at American Education Research Organization.

Marshall, E. C., Duncan, K. D., & Baker, S. M. (1981). The role of withheld information in the training of process plant fault diagnosis. *Ergonomics, 24,* 711-724.

Marshall, E. C., Scanlon, K. E., Shepherd, A., & Duncan, K. D. (1981). Panel diagnosis training for major-hazard continuous-process installations. *The Chemical Engineer.*

Mehle, T. (1980). *Hypothesis generation in an automobile malfunction inference task* (TR 25-2-80). Norman, OK: University of Oklahoma, Decision Process Laboratory.

Mills, R. G., (1971). Probability processing and diagnostic search: 20 alternatives, 500 trials. *Psychonomic Science, 24,* 289-292.

Morris, N. M., & Rouse, W. B. (1984). The effects of type of knowledge upon human problem solving in a process control task. *IEEE Transactions on Systems, Man, and Cybernetics.*

Patrick J., & Stammers, R. (1981). The role of computers in training for problem diagnosis. In J. Rasmussen & W. B. Rouse (Eds), *Human detection and diagnosis of system failures.* New York: Plenum.

Rouse, S. H., & Rouse, W. B. (1982). Cognitive style as a correlate of human problem solving performance in fault diagnosis tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-12,* 649-652.

Rouse, W. B. (1978). Human problem solving performance in a fault diagnosis task. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-8,* 258-271.

Rouse, W. B. (1978). A model of human decision making in a fault diagnosis task. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-8,* 357-361.

Rouse, W. B. (1979). A model of human decision making in fault diagnosis tasks that include feedback and redundancy. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-9,* 237-241.

Rouse, W. B. (1979). Problem solving performance of first semester maintenance trainees in two fault diagnosis tasks. *Human Factors, 21,* 611-618.

Rouse, W. B., & Hunt, R. M. (1984). Human problem solving in fault diagnosis tasks. In W. B. Rouse (Ed.), *Advances in man-machine systems research (Vol. 1)* (pp. 195-222). Greenwich, CT: JAI Press.

Rouse, W. B., & Rouse, S. H. (1979). Measures of complexity of fault diagnosis tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-9,* 720-727.

Rouse, W. B., Rouse, S. H., & Pellegrino, S. J. (1980). A rule-based model of human problem solving performance in fault diagnosis tasks. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-10,* 366-376.

Saltz, E., & Moore, J. V. (1953, February). *A preliminary investigation of troubleshooting* (TR 53-2). Lackland Air Force Base, TX: Human Resources Research Center.

Saupe, J. L. (1954). Troubleshooting electronic equipment: An empirical approach to the identification of certain requirements of a maintenance occupation. (Doctoral dissertation, University of Illinois, 1954.) *Dissertation Abstracts International, 14,* 1966.

Spears, W. (1983). *Processes of skill performance: A foundation for the design and use of training equipment* (NAVTRAEQUIPCEN TR 78-C-0113-4). Orlando, FL: Naval Training Equipment Center.

Steinemann, J. H. (1966, July). *Comparison of performance on analogous simulated and actual troubleshooting tasks* (SRM-67-1). San Diego: U.S. Navy Training Research Laboratory.

Steinemann, J.H., & Hooprich, E.A. (1967). *A performance-oriented electronics technician training program: III. Course evaluation instruments and procedures* (Research Report 68-1). San Diego: Naval Personnel and Training Laboratory. (AD-65-316)

Van Eekhout, J., & Rouse, W. (1981). Human errors in detection, diagnosis and compensation for failures in the engine control room of a supertanker. *IEEE Transactions on Systems, Man, and Cybernetics, 11*(12), 813-816.

Van Matre, N. H., & Steinemann, J. H. (1966, December). *A performance-oriented electronics technician training program: II. Initial fleet follow-up evaluation of graduates* (STB 67-15). San Diego: Naval Personnel Research Activity.

Waltz, E., & Moore, J. V., (1953). *A Preliminary investigation of troubleshooting.* Chanute, IL: Air Force Training Command, Technical Research Center.

APPENDIX A

SUMMARY OF LITERATURE REVIEW

SUMMARY OF LITERATURE REVIEW

Measures Used (Dependent Variables)

Troubleshooting Performance--Quantitative

Any aspect of troubleshooting performance (excluding pencil-and-paper or similar knowledge tests) that has a measurable quantitative dimension.

- Number of problems solved correctly
- Time to solution
- Number of errors
- Costs involved in solving problems
- Time between actions
- Number of correct (relevant, efficient, optimal) actions
- Number of incorrect (illogical, inefficient, redundant) actions
- Ratio of correct to incorrect actions
- Comparison of performance to a statistical optimum
- Rate of learning
- Combination of some of the above measures (cost per error)

Troubleshooting Performance--Qualitative

Any aspect of troubleshooting performance, except for counting the number of occurrences, that cannot be easily quantified with regard to its basic dimensions.

- Overall strategy used
- Strategy used in comparison with an optimum or theoretical method (does performance coincide with "best" or preferred approach)
- Pattern of actions taken
- Types of errors
- "Mental" (cognitive) processes used in solution
- Ratings of the logic of the troubleshooter's actions
- Classification of the strategy used
- Combination(s) of the above measures

Retention

Any measure (usually a quantitative measure) which assesses performance at some period of time after training, initial performance, practice, etc.

Transfer

Some measure (usually quantitative) of performance from one situation (equipment, simulation, classroom training) to a more realistic situation such as actual equipment, performance in operational environment:

- Time saved in reaching a performance criteria
- Number of errors
- Rated overall performance
- Types of errors - Cost savings

Using measures of transfer necessitates a comparison of two or more different environments (types of training, types of equipment) in which the initial training, practice or performance took place.

Good vs. Poor and/or Experienced vs. Inexperienced Troubleshooters

Some studies use existing differences among groups as the dependent measure and other individual characteristics (aptitudes, abilities, system knowledge, cognitive styles, intelligence, experiences) as the independent variables. Supervisor judgments are often used to define good vs. poor. Experience is often defined in terms of number of years on the job vs. students or newly trained individuals. Sometimes a rating, ranking or certification is employed to define experience.

Matrix Format

This section presents a matrix format intended to be used for categorizing and evaluating studies for use of the TAE RDT&E effort. The rationale of "matrixing the research" might be of use to future efforts so the following matrix formats have been included.

Development of matrix formats for categorizing and evaluating the relevant studies was begun but not completed. The general layout format presented below is followed by a series of specific "characteristic" formats related to specific areas of interest.

Information to be placed in the cells of the matrix (factors x measures):

1. The nature of the relationship (direction, size).

2. Special or moderating circumstances of the study with regard to the relationship.

3. Nature of sample(s) - coded.

4. Size of sample.

5. Specification of the actual factor (independent variable) and measure (dependent variable).

6. Reference to any interactions (see interaction factor matrix).

7. Interpretation or theoretical implication of the finding.

8. Reference code.

9. Brief summary of the knowledge/pertinent information (as gathered from the literature) for each factor or measure to be entered into the appropriate cell.

TAE TROUBLESHOOTER CHARACTERISTIC MATRIX FORMAT

Factors	Measures					
Characteristics of the Trouble-shooter	Perf. Quan.	Perf. Qual.	Retention	xfer	Good/bad	Exper/Inexper

1. Specific System Knowledge
2. General Relevant Knowledge (electronics, mechanics, theory)
3. Amount of Training/ Experience
4. Aptitude/Abilities
5. General Intelligence
6. Cognitive Skills/ Styles
7. Other Background Characteristics
8. Combinations of Characteristics

TAE TRAINING CHARACTERISTIC MATRIX FORMAT

Factors		Measures		
Characteristics of Training	Performance Quantitative	Performance Qualitative	Transfer	Retention

1. System specific training
2. General Training, i.e., electronics
3. Overview, theoretical training
4. Combination of two or more of the above or other types of training
5. Strategy training (half split, etc.)
6. Material used in training (actual equipment, manuals, demonstrations, schematics)
7. Abstract vs. system specific

TAE TASK CHARACTERISTIC MATRIX FORMAT

Factor Characteristics of the Task	Measures Performance Quantitative	Performance Qualitative	Retention	Transfer
1. Type of task 2. Abstract vs. system Specific 3. Complexity 4. Type of display 5. Type, amount and timing of feedback 6. Number and types of acceptable paths to solution 7. Real equipment, mock-up simulation				

TAE TASK AIDING MATRIX FORMAT

Factor Characteristics Task Aiding	Measures Performance Quantitative	Performance Qualitative	Transfer	Retention
1. Aids vs. no aids 2. Type of aids (JPA, state tables, tech manuals, schematics, computer recordkeeping) 3. Timing of aiding (during practice/training, initial trials etc.)				

TAE FACTOR INTERACTION MATRIX FORMAT

Factor Interactions Between and Among Two or More Classes of Factors	Measures Performance Quantitative	Performance Qualitative	Transfer	Retention
1. Characteristics of Troubleshooter x Characteristics of the Task				
2. Characteristics of Training x Characteristics of the Task				
3. Characteristics of the Troubleshooter x Characteristics of Training				
4. Characteristics of the Task x Aiding				

APPENDIX B

TEST AND EVALUATION PLAN

TEST AND EVALUATION PLAN[1]

INTRODUCTION

Problem and Background

The Navy faces the problem of being able to objectively measure the technical proficiency of the troubleshooting technician and his ability to contribute to operational readiness. There is no way to evaluate the success of on-board technical training or the effects of hands-on training in the Navy "C" Schools. To address this problem the Navy Personnel Research and Development Center initiated a microcomputer based Troubleshooting Assessment and Enhancement (TAE) research program (originally known as TPEP) in an attempt to measure and diagnose the technical troubleshooting proficiency of Navy personnel. The TAE development effort has resulted in a troubleshooting proficiency demonstration for the high-technology (electronic/digital) maintainer community (NEC ET-1453 for the Naval Modular Automated Communication System (NAVMACS)(V) of the Fleet Satelite Communication (SATCOM) System).

The TAE high-technology demonstration is ready for test and evaluation. The primary goal of the test and evaluation is to determine whether the system provides measures of technical troubleshooting proficiency and diagnostic/remediation capability. The evaluation will provide information on the validity of the TAE diagnostic factors and their relevance in improving remediation approaches and, consequently, troubleshooting proficiency.

To conduct the evaluation it is necessary to develop criterion and diagnostic measures of the personnel participating in the TAE demonstration. Analysis techniques need to be developed to assess the TAE diagnostic factors and testing episodes. Appropriate descriptive, relational and predictive statistical tests need to be performed for data collected during the test and evaluation period in order to provide feedback and recommendations for training system and TAE improvement.

Purpose

The purpose of this appendix is to provide a Test and Evaluation Plan for the TAE NAVMACS/ SATCOM demonstration including:

- Analysis techniques to assess the capability of the TAE diagnostic factors for prescribing remediation training.

- Analysis techniques to validate the ability of the TAE episodes to provide measures of technical troubleshooting performance.

- Analysis techniques for training performance and demographic data to determine TAE's reliability and effectiveness to evaluate troubleshooting proficiency.

- Recommendations for further development and evaluation of TAE.

---

[1]David Dickinson and Sandra Hartley of Instructional Science and Development, Inc. contributed to the preparation of the test and evaluation plan.

METHOD

This section presents the methodology for the TAE NAVMACS(V)/SATCOM Test and Evaluation. It describes the TAE training environment, including the definition of the subject groups and the test administration procedures. The objectives of the test and education plan within this context are defined in terms of reliability and validity. A total of 19 research hypotheses are stated along with the statistical analyses to be performed. Finally, the independent variables for the analysis are defined.

The goal of the TAE effort is to develop a system which can (1) measure troubleshooting performance, and (2) discriminate levels of troubleshooting proficiency. Troubleshooting, within the context of the TAE demonstration, is viewed as part of the corrective maintenance function. When a system is not functioning properly, corrective maintenance must be performed to return the system to an optimum operational state. Troubleshooting is the means by which the dysfunctional component(s) of the system are identified. Once identified, the dysfunctional components can be repaired/replaced. Figure 1 displays this relationship. The focus of the TAE effort is to investigate the ability to troubleshoot by identifying the faulty component.

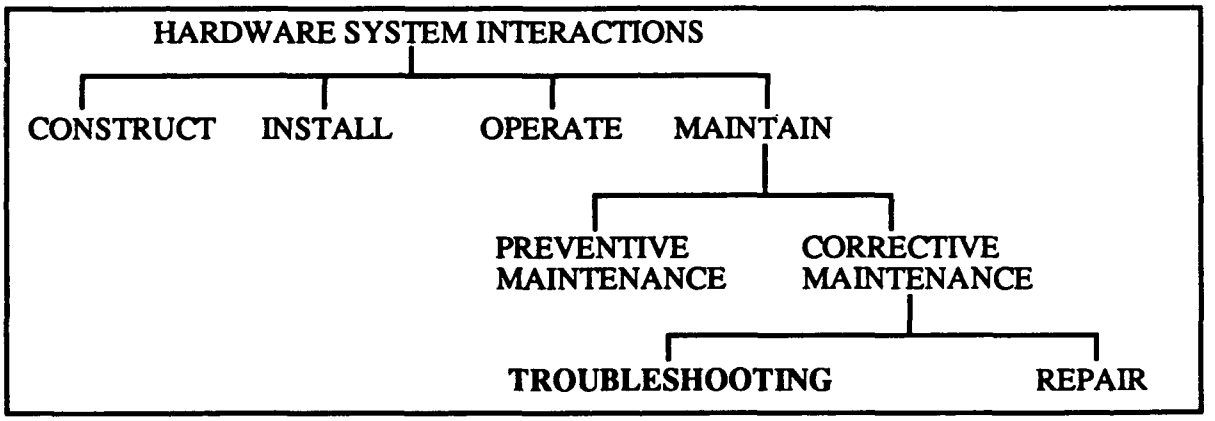

Figure B-1. Hardware activity to troubleshooting.

A review of the literature on troubleshooting techniques and measures, as well as interviews with Navy subject matter experts, resulted in the identification of factors which are relevant in measuring troubleshooting proficiency. By using the ten factors listed below, it should be possible to predict an individual's troubleshooting proficiency level:

1. **Correct Solution** indicates the troubleshooting problem is correctly solved; i.e., the faulty component is identified.

2. **Incorrect Solutions** indicate the Lowest Replaceable Units (LRUs) identified as the faulty component that were not faulty.

3. **Total Time** is the total minutes from login to logout that it took the subject to find the fault.

4. **Test Points** are the total valid reference designator tests.

5. **Proof Points** are test points that positively identify LRUs as faulty. Generally, there will be at least two proof points associated with an LRU, an input and an output point.

6. **Invalid Checks** indicate an inappropriate test was performed at an appropriate test point. For example, a subject measures current where voltage should be checked.

7. **Valid Checks** indicate an appropriate piece of test equipment was used at a test point. For example, a subject measures current where current should be measured.

8. **Redundant Checks** indicate the same test was made at the same test point at sometime during the episode.

9. **Out-of-Bounds** indicate an inappropriate test point was selected. An example would be the selection of a test point that is not reasonably in the area of where the trouble is located.

10. **Illogical Approaches** indicate an inappropriate equipment selection occurred. For example, the subject begins testing on UNIT 7, even though all the symptoms and indications are that the fault is with UNIT 1.

In constructing the TAE test, these troubleshooting factors were integrated into scenarios which require a subject to locate an electronic fault. For the high-tech demonstration, seven *NAVMACS(V)/SATCOM subsystems were used:*

TAE TROUBLESHOOTING HARDWARE

1. TT-624(V)5/UG

2. AN/UYK-20(V)

3. CV-3333/U

4. ON-143(V)/USQ

5. RD-397U

6. AN/USH-26(V)

7. AN/USQ-69(V)

There were no TAE episodes developed for the TSEC/KG-36 due to the sensitivity and classification problems which would have been introduced by developing scenarios based on this equipment.

As shown in the list below, multiple troubleshooting episodes were developed for each NAVMACS(V)/SATCOM subsystem. However, only two TAE episodes for each subsystem will be administered for the test and evaluation. Two of the TAE episodes will be administered as practice scenarios.

TAE TROUBLESHOOTING EPISODES

TT-624(V)5/UG Subsystem

1. Input & Buffer Data Registers[2]
2. Hammer Drivers[3]
3. Paper Feed Control Logic[1]
4. Output Decode
5. Serial Interface Logic

AN/UYK-20(V) Subsystem

1. Channel 16 Interface
2. Micro Channel 15 and IO Oneshot Control[1]
3. Channel 14 Interface[1]
4. Memory Interface
5. Memory Interface

CV-3333/U Subsystem

1. Sample Processor Assembly[1]
2. Sample Data Generator Assembly[1]
3. Spectrum Analyzer No. 2
4. Handset
5. Analyzer and Synthesizer Analog
6. Voicing and Channel Encoder
7. Pitch Analyzer
8. Spectrum Analyzer No. 2
9. Timing and Interface
10. Timing and Self Test

ON-143(V)/USQ Subsystem

1. Level Converter[1]
2. Transmit Sequence Control[1]
3. Relay Card
4. Rec Synchronization
5. Red/Black Interface
6. Red/Black Interface Relay

---

[2]Test and evaluation episodes.
[3]Practice scenarios.

RD-397U Subsystem

1. Punch Enable Signal[1]
2. LDR Signal[1]
3. OD 3 Signal

AN/USH-26(V) Subsystem

1. Formatter A[1]
2. Formatter B[1]
3. Servo/Data
4. Parallel Interface[2]
5. Control

AN/USQ-69(V) Subsystem

1. Maintenance Panel Keyboard
2. Power Supply[1]
3. CRT[1]
4. 2nd, 3rd Page RAM
5. Micro Controller

TAE Subject Groups

The TAE test and evaluation plan is designed to assess the troubleshooting proficiency of three personnel groups within the Navy electronics training and shipboard environments: (1) "C" School students, (2) fleet personnel, and (3) personnel designated as having special assignments. The students are individuals enrolled in "C" School. The fleet personnel are individuals who have graduated from "C" School, hold an NEC ET-1453, and have varying amounts of experience. The special assignment personnel are "C" School Instructors that teach and manage training in the electronics classes, as well as personnel assigned to a Mobile Technical Units (MOTUs).

For experimental purposes, the three different subject groups have been identified as novice, experienced, and expert. Within the novice group there are two sets of individuals who should show the same performance scores: (1) "C" school students and (2) apprentice/inexperienced individuals who have graduated from "C" school and held a NEC ET-1453 rating for less than one year. Individuals with less than one year of experience are considered novices since fleet personnel often do not work in their specific field until after they have been aboard ship for a year. Journeymen/experienced personnel are defined as individuals who hold a NEC ET-1453, are currently assigned to a ship with NAVMACS system ET-1453 billets, and have been working for more than one year in their specific field. Master/experts are defined as individuals that hold a NEC ET-1453 with one year experience working at a MOTU or as a technical representative at a comparable project office. An expert could be a school instructor, a NAVMACS MOTU representative, or project office engineering/technical support military for NAVMACS/SATCOM system at a comparable project office.

---

[1] Test and evaluation episodes.
[2] Practice scenarios.

All TAE subjects must have "C" School training on the NAVMACS(V)/SATCOM subsystems. For students enrolled in "C" School, the test is administered in the last two weeks of training during the "system" phase of the course. The special assignment and experienced fleet personnel groups provide two subject groups with advanced levels of troubleshooting proficiency to compare against the student group with less electronics training and background. The projected number of subjects for the TAE test and evaluation are approximately 100 students, 25 instructors, and 25 fleet personnel.

TAE Test Administration

TAE test administration is completed at the NAVMACS/SATCOM school at Navy Training Center (NTC) in San Diego in a quiet classroom environment. The TAE test is administered on the Zenith 248 microcomputer. The test is completed using technical documentation (hand-held cards) and Maintenance Requirement Cards (MRCs) for the NAVMACS(V) NEC ET-1453 system. All technical documentation and MRCs are within the reach of the subject during testing.

The test administrator assigns the subjects to one of two randomized test scenario formats to protect the TAE study from test order effects. A total of 16 scenarios, including the two practice scenarios, are administered. Each scenario takes about one hour to complete, although there is not a specific time limit. The subjects complete testing on all 14 scenarios in two to three days.

Testing begins with a brief introduction to the TAE study and the technical documentation available. The subjects read and sign a Privacy Act taking the TAE LEARN program. After a short break, the test administrator replaces the LEARN program with the TAE practice and test scenarios. The test administrator is present within the classroom continuously to brief subjects, and set up the programs.

The test administrator starts the subject off by entering his/her social security number. The subject begins with the two practice scenarios to become familiar with the TAE test displays and menus. The TAE testing format begins by displaying fault symptoms. The subject uses a series of menus to review fault symptoms, front panels, maintenance panels, and diagnostic information, to select equipment, to make reference designator tests or replace a Lowest Replaceable Unit (LRU). The subjects goal in the TAE test is to find the faulted LRU as defined by the maintenance philosophy of the system. This is done by selecting the suspected LRU for replacement. It is possible for the fault symptom to logically lead to an LRU that is not the faulted LRU as defined by the scenario. This is indicated as a GOOD FAULT but not the specific faulted LRU. After testing is completed, the subject is given test performance feedback.

Research Objectives

The purpose of the test and evaluation is to provide information on the reliability and validity of the TAE test to discriminate among different levels of troubleshooting proficiency. To establish the objectives of the plan, it is necessary to define the terms reliability, validity, and discrimination within the TAE context. By "discriminate" we mean that if an individual is proficient at troubleshooting, then that individual should score higher on the TAE test than an individual who is less proficient at troubleshooting.

Reliability concerns the problem of errors in measurement. If a measure contains little in the way of measurement error, the measure is said to be reliable. The TAE effort can demonstrate reliability in two ways. One method is by the consistent scoring of the various groups over a number of administrations of the TAE episodes. The expectation is that novices should consistently score lower than the experienced groups. Another method of determining reliability is to have alternative forms of the relevant measure. Each form is administered at separate times, then the correlation between the two forms is computed. The higher the correlations, the more reliable the measure. Since there are two troubleshooting episodes for each subsystem, each episode can be viewed as an alternate form of the same test. The assumption is that there will be a high correlation between an individual's score on the two TAE episodes for the same subsystem.

The general concept of validity revolves around the question, "Does this instrument perform the function it is intended to perform?" More specifically in the case of TAE, the question is "Does TAE discriminate between different levels of troubleshooting proficiency?" It is important to state here that validity is usually a matter of degree as opposed to an all or nothing property. There are three different types of validity depending on the reason the instrument is being used. They are: (1) predictive validity, (2) content validity, and (3) construct validity.

Predictive validity refers to the ability of an instrument to estimate some behavior, known as the criterion. For TAE, the criterion of interest is instrument for predicting the electronics technicians' troubleshooting proficiency in the fleet. However, for the purposes of using TAE as a predictive instrument for troubleshooting proficiency among technicians in the fleet, an emphasis should be placed on the development of a strong criterion measure. An "ultimate" criterion measure of troubleshooting proficiency does not exist. The ultimate criterion would be to obtain the measures that make up the TAE test by having fleet technicians troubleshoot real equipment in a real world situation. Using performance ratings as the criterion measure for technicians in the fleet is problematic since troubleshooting proficiency is only one of many factors that combine to produce job performance ratings within the Navy technical force.

This effort proposes to use fleet subject matter experts (SMEs) and instructor ratings of TAE scoring profiles to construct a troubleshooting proficiency criterion. This measure can be refined over time to produce an ever closer approximation of the ultimate criterion of troubleshooting proficiency. If the concepts of content and construct validity are established, it will be possible to build a strong logical connection between TAE and its ability to predict troubleshooting proficiency among electronics technicians in the fleet.

TAE can be empirically validated for use as a predictive instrument for success in the "C" school program by using the various "C" school test scores as criterion measures. The assumption is the TAE test given prior to the "C" school performance tests can accurately predict the subsequent "C" school performance test scores.

Content validity is concerned with the question, "Does the instrument adequately sample a particular domain?" If troubleshooting proficiency is viewed as being composed of skills and knowledges, then determining the content validity of TAE is a legitimate task. Content validity is more an appeal to the representativeness of the content items of the test and the manner in which it is presented as opposed to empirical validity (Nunnally & Durham, 1975). Although empirical findings can lead to more confidence in confirming content validity.

For a representative sampling of content items, a detailed outline containing the knowledge domain must be developed. The TAE staff outlined and reviewed a large domain of information pertaining to troubleshooting. In addition, 750 Subject Matter Experts (SMEs) were surveyed to establish the troubleshooting factors. When evaluating and assessing the content validity of the outlined troubleshooting domain and troubleshooting factors, the important questions/problems/ issues within the troubleshooting domain appear to have been represented. Of course, any content outline has limitations. Content validity is affected by the popular theory or subject emphasis within that group for that time period, thus content validity is not consistent through time. Yet, for troubleshooting research pertaining to the Navy's unique requirements, the troubleshooting subject matter reviewed appears to cover the pertinent knowledge domain.

The TAE test must be in a constructed logical and sensible manner. This is a relatively straightforward task when constructing test items for concrete, easily observable content. On the other hand, developing a test sensibly for content domain that is abstract and difficult to define (as is troubleshooting) is challenging. The big question is whether or not the content has been adequately sampled and developed into test items (Nunnally & Durham, 1975). It appears that the domain of troubleshooting is well covered within the TAE test and reflects actual equipment faults that technicians may encounter in the fleet. Yet, troubleshooting is considered an area that is difficult to test sensibly since it measures areas which are abstract and require extensive skill. For areas that are abstract and complex, there is debate over what is the best method for testing understanding of the area--i.e., questions, problems, work samp' :s. TAE testing involves troubleshooting problems delivered by means of computerized scenari.s. It is yet to be determined if this is the best way to sample the subject's domain of knowledge and skill in troubleshooting.

Troubleshooting proficiency can be viewed as a psychological trait measured by performance on the TAE test. It then becomes important to validate the use of TAE as a test of this psychological trait. This type of validation involves construct validity. Another way of saying this is that variance in TAE scores among subjects is due primarily to the construct of troubleshooting proficiency. According to Nunnally and Durham (1975), construct domains differ in how (1) large or small or (2) tightly or loosely defined the observable variables are. Larger domains may encompass a large amount of variables that are difficult to define and the variable definitions are not clear. With a large domain like troubleshooting, it is essential that all subject material and factors are clearly defined.

The three essentials for creating and validating a construct measure are to: (1) outline the domain of observables, (2) find out which and how much the different observables relate with each other or are affected similarly by experimental treatments, and (3) find out whether one or more of the observable variables measure the construct (Nunnally & Durham, 1975).

The domain of observables has been defined in terms of the troubleshooting factors, such as finding the solution, valid checks, number of tests made, etc. Through empirical tests, it is possible to determine which of the troubleshooting observables correlate with each other or are affected alike by TAE treatments. The measures that respond similarly and consistently for the different treatments are seen as holding the most construct validity. Within the TAE test, the troubleshooting factors will be investigated to determine whether they respond in a similar and consistent manner to the different TAE episodes. The third measure of construct validity is whether the measure(s) of the construct behave as expected. For example, students (novices) would be expected to score lower on the TAE test than more experienced fleet personnel. Many of the hypotheses that have

been generated in this effort relate to the construct validity of TAE. Taken together, they will provide support for the validation of TAE as a measurement of troubleshooting proficiency.

Research Hypotheses and Analyses

The TAE evaluation plan is designed to test a total of 19 hypotheses. Data will be collected and analyzed for an unequal number of subjects: 100 students, 25 experts, and 25 experienced fleet personnel. A composite score will be obtained for the 14 total TAE test scenarios. The data will be subjected to a one-way analysis of variance (ANOVA) design using groups as the experimental factor. The hypotheses are organized into seven categories: (1) experience, (2) electronics knowledge, (3) electronics performance proficiency, (4) difficulty level, (5) time, (6) complex test equipment, and (7) ranking. The hypotheses in each category are stated and then the statistical techniques to analyze the hypotheses in each category are described in the following paragraphs.

Experience Hypotheses (1 through 3)

It has been noted in previous research that experience is positively correlated with TS proficiency. According to Vineberg (1968), both field experienced mechanics and trainees exhibited differences in TS proficiency attributed to experience (Morris & Rouse, 1985). Therefore:

1. Instructors will score significantly higher than students (novices) on the TAE test than students (novices).

2. Experienced fleet personnel will score significantly higher on the TAE test than students (novices).

3. Subjects with a longer time in the electronics rate will score significantly higher on the TAE test than subjects with less time in that rate.

Hypotheses 1 and 2 will be tested using one-way ANOVAs that compare the group means of the TAE test scores for: (1) instructors vs. students, (2) experienced fleet personnel vs. students, (3) instructors vs. experienced fleet personnel. Results should indicate that both instructors and fleet personnel score significantly higher on the TAE test than students (novices). These results will also support the validity of the TAE test to measure TS proficiency since both instructors and experienced fleet personnel are hypothesized to be more proficient troubleshooters. There is no expected significant difference between instructors and experienced fleet personnel. Hypothesis 3 will be tested with a Pearson correlation with the expectation that there will be a positive relationship between time in the electronics rate and TAE test score.

Electronics Knowledge (Hypotheses 4 through 6)

It appears that subject TS performance indicators (such as "C" school scores and appropriate ASVAB scores) will be positively correlated with TS ability on the TAE test. Therefore:

4. Students (novices) with higher academic "C" school final scores will score higher on the TAE test than students (novices) with lower scores.

5. Students (novices) with higher academic "C" school subsystem scores will score higher on the TAE subsystems than students (novices) with lower "C" school subsystem scores.

6. Subjects with higher appropriate Armed Services Vocational Aptitude Battery (ASVAB), (i.e., EI, Electronics Information) and ET selection criteria scores will score higher on TAE test than students with lower ASVAB EI and selection scores.

Hypotheses 4, 5, and 6 will be tested using a Pearson correlation. The expected results should indicate a positive correlation between final grades in "C" school and the TAE test score, as well the "C" school subsystem scores and the TAE subsystem scores. A positive relationship is expected between the ASVAB score and the TAE test score. Hypothesis 5 can be further elaborated by using scores on paper and pencil knowledge-based "quizzes" for one set of correlations and scores on performance tests as another set of correlations. The set of TAE test-performance score correlations will support both the construct and concurrent validity of the TAE test. This set of correlations will also support the reliability of the TAE test since the subjects will be troubleshooting the same subsystems, possibly even the same piece of equipment.

Electronics Performance Proficiency (Hypotheses 7 through 11)

A number of previous reports have indicated that the technical knowledge or practical job knowledge is related to TS performance (Morris & Rouse, 1985). It also seems reasonable to assume that good troubleshooters will be more correct in their choices of test points to check and test equipment to use. Therefore:

7. Subjects with a higher level of TS proficiency will make fewer invalid checks than less proficient subjects.

8. Subjects with a higher level of TS proficiency will make fewer illogical approaches than less proficient subjects.

9. Subjects with a higher level of TS proficiency will have fewer incorrect solutions than less proficient subjects.

10. Subjects with a higher level of TS proficiency will make fewer redundant checks than less proficient subjects.

11. Subjects with a higher level of TS proficiency will test significantly more proof points than less proficient subjects.

Hypotheses 7 through 11 will be examined through correlational analyses. In hypotheses 7 through 10, there is an expectation to find an inverse relationship between the TAE test score and the specific TAE factor. In hypothesis 11, a positive correlation is expected between the TAE test score and proof points. Each of these hypotheses will compare the TAE test score and a measure contained within the TAE test score to compute correlation coefficients. If the scoring formula contains specific TAE factors, partial correlations will be used. Hypotheses 7 through 11 will also support the construct validity of TS proficiency.

Difficulty Level (Hypotheses 12 through 15)

It seems reasonable to assume that increasing TS task difficulty will increase the time expended in finding the solution. The length of time to solution would also be affected by the subject's TS proficiency level. Therefore:

12. The more difficult the scenario is, the longer will be the average time to find the solution across subjects.

13. On scenarios of equal difficulty, subjects with a higher level of TS proficiency will take significantly less time than less proficient subjects in finding the solutions.

14. The more difficult the scenario is, the less time the instructors will take to find the TAE test solutions when compared to the students (novices).

15. The more difficult the scenario is, the less time the experienced fleet personnel will take to find the TAE test solutions when compared to the students (novices).

It has not yet been decided how to determine the episode difficulty level. Difficulty level could be determined from a Pearson correlation taken from a combination of factors; i.e., total time, total steps and LRUs replaced incorrectly. An additional measure for difficulty level could be for the subjects to rank the difficulty level of the episodes after taking the TAE tests. One broad measure of difficulty level could be whether or not the scenario contains Reference Designator Tests (RDTs) as compared to Diagnostic Tests (DTs) with the assumption being that RDTs involve a more complex decision process than DTs.

Hypothesis 12 can be tested using a Pearson correlation with the expectation of finding a positive correlation between level of scenario difficulty and total time in finding solution. Hypothesis 13 assumes that difficulty level of the scenarios can be determined. It can be tested using a correlational analysis with the expectation of finding an inverse relationship between the TAE test score and time. Again, if the TAE scoring formula contains time as a factor, partial correlations will be used.

Hypotheses 14 and 15 can be tested using 1-way ANOVAs. Expected results should show that both instructors and experienced fleet personnel will find the test solutions in significantly less time than the students (novices). There is no expected significant difference between the instructors and experienced fleet personnel.

Time Specific (Hypotheses 16 and 17)

It was noted during discussions with subject matter experts that good troubleshooters often will take a longer period of time to make the first test of equipment. This observation seems related to previous research concerning cognitive styles and TS where it was noted that subjects with a reflective vs. an impulsive cognitive style made fewer errors in TS tasks (Morris & Rouse, 1985). It may be that a good troubleshooter begins by surveying the state of the equipment to generate hypotheses about the possible fault, uses the test to collect information, and then takes a longer

amount of time to integrate the information discovered to generate more hypotheses about the problem. Therefore:

16. In general, subjects with the higher level of TS proficiency will take a significantly longer time than less proficient subjects before making the first test.

17. In general, subjects with a higher level of TS proficiency will make significantly fewer tests than less proficient subjects.

Hypothesis 16 will be tested using a TAE test score and length of time to first test. The rationale behind this hypothesis is that subjects who have a high level of TS proficiency will use more time to generate hypotheses about what the possible fault is before they make the first test.

Hypothesis 17 will be tested using a Pearson correlation with the expected results of an inverse relationship (i.e., the higher the TAE test score is the fewer tests made).

Complex Test Equipment (Hypothesis 18)

Previous research has noted that good troubleshooters tended to make more difficult checks than poor troubleshooters (Saltz & Moore, 1953). It would seem reasonable to state that good troubleshooters will use more complex test equipment. Therefore:

18. Subjects with a higher level of TS proficiency will make significantly more tests using an oscilloscope than will less proficient subjects.

Ranking (Hypothesis 19)

Finally, it seems reasonable to assume that if the TAE test reflects an individual's TS proficiency, there will be a positive relationship between the composite TAE score and instructor rankings of the score profiles.

19. The higher the subject's TAE score is, the higher the subject will be ranked in terms of TS proficiency by instructors/work center supervisors.

Definition of Variables

The dependent variable is the TAE test score, which is a composite score made up of the subject's performance on the following factors:

1. **Solution** - Bivariate distribution; subject correctly/incorrectly solved the problem.

2. **Proof Points** - Integer; the total number of proof points tested. A proof point is a test point that positively identifies an LRU as faulted.

3. **Time**

Total Time - total number of minutes from login to logout.

Time before First Test, whether a Reference Designator test or a Diagnostic test.

4. **Test Points** - Integer; total number valid reference designator tests.

5. **Checks** - All integers.

Invalid Checks - total number of invalid checks. An invalid check is when a subject uses an inappropriate piece of test equipment at a test point.

Valid Checks - total number of valid (good) checks. A valid check is when a subject uses an appropriate piece of test equipment at a test point.

Redundant Checks - total number of same test types made consecutively at the same test point.

6. **Illogical Approaches** - Integer; total number of times an illogical approach is used. An illogical approach indicates an inappropriate equipment selection occurred.

7. **Incorrect Solutions** - Integer; total number of times the subject replaced a Lowest Replaceable Unit (LRU) incorrectly when it was not the fault.

8. **Out-of-Bounds** - Integer; total number of times an out-of-bounds test was made.

Independent variables include performance, academic, and demographic variables. Performance factors include the various "C" school performance tests scores, while academic variables include the "C" school subsystem scores, ASVAB EI (Electronics Information) and other appropriate ASVAB scores, scenario difficulty level, three subject groups (students, instructors, and fleet personnel), and the instructor/work center supervisor ratings. Demographic variables include age, rate, primary and secondary NECs held, etc.

RESULTS

Development of the TAE NAVMACS(V)/SATCOM Test and Evaluation Plan resulted in a number of items necessary to analyze the TAE data to be collected. These items include:

1. The development of formats for the performance data and demographic data to be collected.

2. The preparation of guidelines to be used by personnel whose familiarity with statistics and with statistical procedures is somewhat limited.

3. The development of job aids for using the MICROSTAT[3] statistical package.

In general, analyses will be performed by using one or both of the databases to create analyses files through the MICROSTAT statistical package. These separate analyses files will then be used for the actual data analysis. Guidelines for performing the statistical tests will be provided to ensure standardized techniques are used.

---

[3]Identification of specific equipment and software is for documentation only and does not imply any endorsement.

Database Formats

Two separate databases will be developed. One database will contain the performance information collected when a subject takes a TAE episode. The other database will contain demographic information such as age, gender, NECs held on each subject.

Data Format for Subject Performance Database

Each case contains all the performance data for one subject. Each student performs 16 episodes, including two practice episodes and the 14 test scenarios. Data for the practice episodes will be included for completeness, but they will not be used for analysis. The episodes in the performance database file will always appear in the following order for each subject:

1. USH26, episode 1
2. USH26, episode 2
3. USQ69, episode 2
4. USQ69, episode 3
5. UYK20, episode 2
6. UYK20, episode 3
7. CV3333, episode 1
8. CV3333, episode 2
9. ON143, episode 2
10. ON143, episode 3
11. RD397, episode 1
12. RD397, episode 2
13. TT624, episode 1
14. TT624, episode 3
15. USH26, episode 4 "practice"
16. TT624, episode 2 "practice"

The data file produced by Editview will contain a total of 673 variables for each case (subject). There are 42 variables for each of the 16 episodes plus the first variable (V1) for the subject's social security number. If a subject does not perform any given episode, the Editview program will output the Equipment number and Episode number, followed by 40 "MISSING" variables, in order to fill out the space normally taken up by the episode data. Table B-1 describes the variables for each episode, while Table B-2 identifies the variables for each case.

Data Format for Subject Demographic Database

Demographic data files will be set up for each of the three subject groups: students, instructors,and fleet personnel. The three files will be set up to contain demographic data for 100 students, 25 instructors, and 25 fleet personnel. Each student case will contain 52 variables (D1 through D52). The instructor and fleet personnel cases may only contain variables D1 through D22 since the "C" school data may not be available for these subjects. Table B-3 describes the demographic data file variables.

Table B-1

Description of Variables for a TAE Episode

Variable Name	Contents of Variable
V1	Subject's Social Security Number
V2	Equipment (hardware subsystem) number (1 = USH26)
V3	Episode number (1)
V4	Found Solution (1 = Yes, 0 = No)
V5	Number of Test Points
V6	Number of Out of Bounds tests
V7	Number of Valid Checks
V8	Number of Invalid Checks
V9	Number of Redundant Checks
V10	Number of Proof Points subject tested
V11	Total number of Proof Points in the episode
V12	Percentage of Proof Points tested: (V10 % V11) * 100, rounded to a whole number
V13	Total Time spent on the episode (in Minutes)
V14	TBD
V15	Number of Equipment Selection events
V16	Number of Front Panel events
V17	Number of Maintenance Panel events
V18	Number of Fallback test events
V19	Number of Reference Designator test events
V20	Number of Replace LRU events
V21	Number of Review Symptoms events
V22	TBD
V23	Number of Diagnostic Test events
V24	Number of Load Operational Program events
V25	Number of Step Procedure events
V26	Number of Revision events
V27	Number of INCORRECT Replace LRU events
V28	Number of GOOD FAULT REPLACEMENT Replace LRU events
V29	Time to first Reference Designator Test (in minutes)
V30	Time to first Diagnostic Test (in minutes)
V31	Total number of steps taken in the episode: ALL events, (even "login" and "logout") except "revision" events, which are created when an instructor edits episode data.
V32	Number of Waveform tests performed
V33	Number of Voltage tests performed
V34	Number of Read Meter tests performed
V35	Number of Logic tests performed
V36	Number of Current tests performed
V37	Number of Frequency tests performed
V38	Number of Continuity tests performed
V39	Number of Adjustment tests performed
V40	Final Score of the episode
V41, V42, V43	TBD -- these are for possible future expansion

TBD = to be determined.

Table B-2

Description of TAE Variables for Each Case

USH26		USQ69		UYK20		CV3333		ON143		RD397		TT624		Practice		
V2	44	86	128	170	212	254	296	338	380	422	464	506	548	590	632	Equip (subsys)
V3	45	87	129	171	213	255	297	339	381	423	465	507	549	591	633	Episode number
V4	46	88	130	172	214	256	298	340	382	424	466	508	550	592	634	Found Solution
V5	47	89	131	173	215	257	299	341	383	425	467	509	551	593	635	Test Points
V6	48	90	132	174	216	258	300	342	384	426	468	510	552	594	636	Out-of-Bounds
V7	49	91	133	175	217	259	301	343	385	427	469	511	553	595	637	Valid Checks
V8	50	92	134	176	218	260	302	344	386	428	470	512	54	596	638	Invalid Checks
V9	51	93	135	177	219	261	303	345	387	429	471	513	555	597	639	Redun. Checks
V10	52	94	136	178	220	262	304	346	388	430	472	514	556	598	640	Proof Points
V11	53	95	137	179	221	263	305	347	389	431	473	515	557	599	641	Tot. PPs/Epis.
V12	54	96	138	180	222	264	306	348	390	432	474	516	558	600	642	Percentage PPs
V13	55	97	139	181	223	265	307	349	391	433	475	517	559	601	643	Total Time
V14	56	98	140	182	224	266	308	350	392	434	476	518	560	602	644	TBD
V15	57	99	141	183	225	267	309	351	393	435	477	519	561	603	645	Eq.Sel.Events
V16	58	100	142	184	226	268	310	352	394	436	478	520	562	604	646	Front Panel Ev.
V17	59	101	143	185	227	269	311	353	395	437	479	521	563	605	647	Maint Panel Ev.
V18	60	102	144	186	228	270	312	354	396	438	480	522	564	606	648	Fallback Events
V19	61	103	145	187	229	271	313	355	397	439	481	523	565	607	649	Ref Desig Tests
V20	62	104	146	188	230	272	314	356	398	440	482	524	566	608	650	Replace LRU Ev.
V21	63	105	147	189	231	273	315	357	399	441	483	525	567	609	651	Review Symp Ev.
V22	64	106	148	190	232	274	316	358	400	442	484	526	568	610	652	TBD
V23	65	107	149	191	233	275	317	359	401	443	485	527	569	611	653	Diag Test Ev
V24	66	108	150	192	234	276	318	360	402	444	486	528	570	612	654	Load Op Prgm Ev.
V25	67	109	151	193	235	277	319	361	403	445	487	529	571	613	655	Step Proced Ev.
V26	68	110	152	194	236	278	320	362	404	446	488	530	572	614	656	Revision Events
V27	69	111	153	195	237	279	321	363	405	447	489	531	573	615	657	INC Rep LRU Ev.
V28	70	112	154	196	238	280	322	364	406	448	490	532	574	616	658	GdFaultRelLRU Ev
V29	71	113	155	197	239	281	323	365	407	449	491	533	575	617	659	Ti 1st RefDesTst
V30	72	114	156	198	240	282	324	366	408	450	492	534	576	618	660	Ti 1st Diag Test
V31	73	115	157	199	241	283	325	367	409	451	493	535	577	619	661	Total Steps
V32	74	116	158	200	242	284	326	368	410	452	494	536	578	620	662	Waveform Tests
V33	75	117	159	201	243	285	327	369	411	453	495	537	579	621	663	Voltage Tests
V34	76	118	160	202	244	286	328	370	412	454	496	538	580	622	664	Read Meter Tests
V35	77	119	161	203	245	287	329	371	413	455	497	539	581	623	665	Logic Tests
V36	78	120	162	204	246	288	330	372	414	456	498	540	582	624	666	Current Tests
V37	79	121	163	205	247	289	331	373	415	457	499	541	583	625	667	Frequency Tests
V38	80	122	164	206	248	290	332	374	416	458	500	542	584	626	668	Continuity Tests
V39	81	123	165	207	249	291	333	375	417	459	501	543	585	627	669	Adjust. Tests
V40	82	124	166	208	250	292	334	376	418	460	502	544	586	628	670	Final Score
V41	83	125	167	209	251	293	335	377	419	461	503	545	587	629	671	TBD
V42	84	126	168	210	252	294	336	378	420	462	504	546	588	630	672	TBD
V43	85	127	169	211	253	295	337	379	421	463	505	547	589	631	673	TBD

Table B-3

Description of Demographic Variables for a TAE Case

Variable Name	Contents of Variable
D1	Social Security Number
D2	Age
D3	Gender
D4	Rate
D5	Date of Rate
D6	Time in Rate
D7	Time in Service
D8	Primary NEC
D9	Secondary NEC
D10	Secondary NEC
(The following are ASVAB Scores.)	
D11	GI
D12	NO
D13	AD
D14	WK
D15	AR
D16	SP
D17	MK
D18	EI
D19	MC
D20	GS
D21	SI
D22	AI
(The remaining variables are "C" School data.)	
D23	Class #
D24	Graduation Date
(The next four entries are for the USH26/USQ69/RD397/TT624 equipment.)	
D25	Quiz Average
D26	PT's Percentage
D27	Test Score
D28	Area Total
(The next four entries are for the UYK20 equipment.)	
D29	Quiz Average
D30	PT's Percentage
D31	Test Score
D32	Area Total

Table B-3 (Continued)

Variable Name	Contents of Variable
(The next four entries are for the CV3333 equipment.)	
D33	Quiz Average
D34	PT's Percentage
D35	Test Score
D36	Area Total
(The next four entries are for the ON143 equipment.)	
D37	Quiz Average
D38	PT's Percentage
D39	Test Score
D40	Area Total
(The next four entries are for the KG36 equipment.)	
D41	Quiz Average
D42	PTs Percentage
D43	Test Score
D44	Area Total
(The next four entries are for systems equipment.)	
D45	Quiz Average
D46	PT's Percentage
D47	Test Score
D48	Area Total
(The next three entries are "C" School totals.)	
D49	Class Standing
D50	Total Number in Class
D51	Final Score
D52	Student Rankings by Instructor

Guidelines for Statistical Analyses

Guidelines for the statistical analyses to be performed during the test and evaluation were prepared for use by personnel whose statistical background is somewhat limited. "Somewhat limited" means they are familiar with the concepts of mean, median, mode, variance, and standard deviation. Guidelines were developed for the following:

1. Preliminary evaluation of the data set .

2. Standard multiple regression analysis.

3. One-way between-subjects analysis of variance (ANOVA).

4. Pearson correlation formula.

5. Point Biserial correlation formula.

While the MICROSTAT statistical package has options for most of the analyses mentioned above, there is no option for a Point Biserial correlation. Some of the separate steps involved in the Point Biserial can be performed in MICROSTAT but it also requires the use of a calculator. Since the Point Biserial is not a menu option, that section is more specific in terms of the individual steps involved in its computation than the other analyses which are menu options within MICROSTAT.

Guidelines for Preliminary Evaluation of Data

There are six checks that should be performed on the final data set prior to any data analysis. This preliminary evaluation will be used to "clean up" the data set and to pinpoint any violations of the assumptions that underlie many of the statistical procedures. These violations may lead you to perform variable transformations or to limit your conclusions resulting from the data analysis. A description of each check and the steps that need to be performed are provided below.

1. Inspect univariate descriptive statistics for accuracy of input.

A. Check for out-of-range values.

(1) Use MICROSTAT Descriptive Statistics function to identify minimum and maximum.

(2) Check to see that each minimum and maximum is within the possible range of values.

B. Check for plausible means and dispersions.

(1) Use MICROSTAT Descriptive Statistics to obtain means and standard deviations.

(2) Check to see if means and standard deviations appear reasonable.

2. Evaluate number and distribution of missing data; determine best approach on how to deal with missing data.There are four basic approaches to the problem of missing data:

A. Treat missing data as data. The assumption is that the data are missing due to some behavior of the subject such as refusing to respond to a question. This is probably not a relevant approach with TAE.

B. Delete cases or variables. This assumes that the cases or variables are randomly distributed throughout the data set. If this is not the case, then the data sample becomes distorted by dropping cases or variables. For TAE, dropping those variables that have values of 0 across subjects for all scenarios (such as current tests or continuity tests) will reduce the size of the data set without the loss of significant data.

C. Estimate missing data by using the mean of the sample. Inserting the mean value of a variable is often the best approach when there is no other information to help you determine its value. One possible disadvantage to this approach is that it may lead to a lower correlation between this variable and other variables if there are numerous missing values. This is appropriate in a situation where there are few missing values and is the most conservative of the various approaches. This is probably the most appropriate for the TAE effort.

D. Use regression to predict missing values. It is possible to construct a regression equation to estimate missing values by using variables that are available as the intervening variables (IVs) to predict the missing value. This approach requires that the variables used as the IVs be correlated with the variable to be predicted. It may also lead to an "overfitting" of the data, meaning that the results will apply only to the sample and not be generalizable to the population.

3. Identify and deal with outliers.

A. Univariate outliers.

(1) Dichotomous variables. Dichotomous variables that have splits of 90%-10% may lead to misleading correlations and should be evaluated.

(2) Continuous variables.

a. Scores need to be transformed to Z-scores. The mean and standard deviations should first be computed using MICROSTAT Descriptive Statistics.

b. Select the Recode/Transform/Select option in MICROSTAT.

c. Select the Z-TRANS function and use the mean and standard deviation to compute Z-scores.

d. Consider as an outlier any value which has a Z-score greater than +/-3.

B. Once outliers have been identified, there are several courses of action to consider.

(1) Check to see if the data were input correctly.

(2) Delete outliers with great care. If the specific circumstances during testing that are responsible for the outlier (such as machine failure) can be identified, then this is an appropriate method. This applies to both continuous and dichotomous variables. Obviously, if the information contained in the "outlying" dichotomous variable is of interest, this method is impractical.

(3) Apply an appropriate data transformation so that outliers are moved closer to the mean.

4. **Identify and deal with skewness.** Skewness refers to the shape of the distribution. In general, the various TAE factors will exhibit some form of skewness.

A. Locate skewed variables. Skewness refers to the observation that values for a variable tend to pile up at one end or the other of the distribution. This situation can be detected by using the Frequency Distribution function in MICROSTAT.

B. Transform skewed variables (if desirable). Generally skewness is not a problem in most statistical analyses especially if there is a large sample size. Various data transformations are available through MICROSTAT if the data are grossly skewed. Generally, you would use a log type transformation to reduce a positively skewed distribution and an exponential type transformation for a negatively skewed distribution.

C. Check results of transformation. This would also be done with the Frequency Distribution function in MICROSTAT.

5. **Identify and deal with nonlinearity and heteroscedasticity.**

A. Since correlation coefficients measure the linear relationship between two variables, nonlinear relationships will result in artificially low values. Nonlinearity can be detected by using the Scatterplot function in MICROSTAT to produce bivariate scatterplots. Generally, only gross departures from linearity need be rectified. This can be done through various data transformation available in MICROSTAT.

B. Heteroscedasticity refers to the situation where variability in scores on one variable differs across values on the other variable. Since correlational analysis assumes the same variability across values for both variables (homoscedasticity), heteroscedasticity will lead to an underestimation of the true relationship between two variables. It may be possible to reduce heteroscedasticity through data transformations.

6. **Evaluate variables for multicollinearity.** Multicollinearity refers to the fact that two variables are almost perfectly correlated. Generally, if a variable is almost perfectly correlated with one or another (correlation of .99 or greater) of the variables and share the same pattern of correlations with other variables, it can be dropped from the analysis. Logical considerations should determine which of the two to actually drop.

Guidelines for Standard Multiple Regression

The following guidelines address some of the major issues and analysis techniques for standard multiple regression. TAE will primarily be concerned with using a regression equation for purposes of prediction.

1. Issues

A. Number of cases and variables. Generally the issue here is the ratio of cases to variables. A minimum number of cases to variables might be 15 or 20 to 1. Some considerations are:

(1) Skewness of DV - The more skewed the DV, the more cases you need (assuming no data transformation is done).

(2) Effect size - More cases are needed to demonstrate a small effect size.

(3) Measurement error - The greater the measurement error in the measuring instrument the more cases are needed to demonstrate.

B. Outliers. Evidence of outliers calls for the same approach taken in the preliminary evaluation of data. The data should be checked carefully for incorrect input. Also, extreme cases should be deleted or the data set transformed.

C. Multicollinearity. Again, variables that are either perfectly or nearly perfectly correlated should be dealt with by eliminating one of the variables.

2. Major Analyses

A. Multiple R, F ratio. Once a multiple R has been obtained through MICROSTAT Regression Analysis function, it needs to be determined if it is statistically significant. This can be done by checking the Probability (PROB.) value given through the Regression Analysis function. A value of .05 or less indicates statistical significance.

An example from an actual MICROSTAT output is shown below. Since the value shown here (underneath PROB.) is less than .05, this is a statistically significant Multiple R.

ADJUSTED R SQUARED	=	.5998
R SQUARED	=	.6172
MULTIPLE R	=	**.7856**

ANALYSIS OF VARIANCE TABLE

SOURCE	SUM OF SQUARES	D.F.	MEAN SQUARE	F RATIO	PROB.
Regression	87549.4019	3	19001.1340	18.9296	.004
Residual	54303.1195	66	1003.7745		
Total	141852.5214	69			

B. Adjusted R SQUARED, overall proportion of variance accounted for. This value, which is also part of the Regression Analysis output, is an adjustment in the R SQUARED which takes into account the possible overestimation of the R SQUAREDThis value indicates what proportion of the variance in the dependent variable is accounted for by the variables in the prediction equation. R SQUARED is simply the MULTIPLE R squared.

ADJUSTED R SQUARED	**= .5998**
R SQUARED	= .6172
MULTIPLE R	= .7856

C. Significance of regression coefficients. Whether or not the regression coefficient is statistically significant can be determined by checking the output from the MICROSTAT Regression Analysis function. Each variable entered into the regression equation will have a PROB.value associated with it.

This value should be less than .05 to indicate statistical significance.

VAR.	REGRESSION COEFFICIENT	STD. ERROR	T(DF= 66)	PROB.	PARTIAL r^2
var1	-.8024	.2433	-3.299	.00157	.1415
var2	-2.7583	.7536	-3.660	.00050	.1687
var3	-1.6825	.6093	-2.761	.00745	.1036

D. Squared semipartial correlations. This value represents the unique contribution of an IV to the squared multiple correlation (R^2). This value is part of the output of the Regression Analysis function.

VAR.	REGRESSION COEFFICIENT	STD. EFFORT	T(DF= 66)	PROB.	PARTIAL r^2
var1	-.8024	.2433	-3.299	.00157	**.1415**
var2	-2.7583	.7536	-3.660	.00050	**.1687**
var3	-1.6825	.6093	-2.761	.00745	**.1036**

Guidelines for One-Way Between-Subjects Analysis of Variance

The following guidelines concern some of the issues involved in using a one-way between-subjects ANOVA to determine if there are significant differences between the three subject groups involved in the TAE study. The ANOVA allows us to explore the variability between the three subject groups. Within the TAE study, there is one dependent variable, the TAE score. The one independent variable is subjects, with the three levels representing the three subject groups.

1. Assumptions and Limitations. The statistical model underlying the ANOVA assumes that:

   A. Subject groups were drawn from populations that are normally distributed.

   B. The subject group variance is homogeneous.

   C. Subjects have been randomly and independently assigned to the different treatment groups.

These assumptions are conservative. The test of significance for the ANOVA (F-test) is very robust. Unless these assumptions are grossly violated or the probability value is on the borderline of significance then the results can be considered valid.

2. Major Analyses

A. F Value. The F ratio from the ANOVA allows us to find out if there is a significant difference between the means of the three TAE subject groups. The MICROSTAT One-Way Analysis of Variance package will display the means and number of subjects for each of the groups. The ANOVA package then displays a summary table with the F value and PROB. (probability). An example from an actual output is shown below.

ONE-WAY ANOVA

GROUP	MEAN	N
1	418.014	35
2	427.300	35
GRAND MEAN	422.657	70

VARIABLE 4: finscr

Source	Sum of Squares	D.F.	Mean Square	F Ratio	PROB.
BETWEEN	1508.929	1	1508.929	.731	.3955
WITHIN	140343.593	68	2063.876		
TOTAL	141852.521	69			

B. Significance of the F Value

(1) The PROB (probability) within the summary table indicates the level at which the F value listed in the summary table is significant. An F value probability of .0500 or less is interpreted as significant and the null hypothesis is rejected.

(2) If the F ratio is significant, then we go one step further to perform a specific-comparison test to find out if the specific means significantly differ from each other.

Guidelines for Pearson r Correlation

The following guidelines concern the use of the Pearson r correlation to determine if a relationship exists between two variables, its direction, and its strength.

1. Issues

A. The Pearson r investigates the strength of the relationship between two variables. There must be pairs of measurements on two variables to conduct a Pearson r correlation. Each subject will have a set of values, i.e., score or rank variables.

B. The Pearson r correlation coefficient number indicates the magnitude and the sign of the coefficient indicates the direction of the relationship. The Pearson r correlation ranges from -1 as a negative (inverse) relationship, 0 as no relationship and +1 as a positive relationship between the variables. The closer to 1 or -1 is the stronger the relationship between the variables.

2. Major Analyses

A. You can obtain the Pearson r correlation coefficient between two variables using the MICROSTAT Correlation Matrix analysis.

B. Significance of the Pearson r correlation. To determine the significance of the Pearson correlation coefficient. On the MICROSTAT correlation matrix display, look for the critical values listed last. The critical values for a 1-tail and 2-tail tests at the .05 significance level are listed. If the r obtained in the correlation matrix is greater than the r critical value, then your Pearson r is significant.

In the example below, the critical value is a +/- .23502. The correlation between the two variables, tsteps and finscr, is -.74295. Since -.74295 is greater than -.23502, there is a significant negative correlation between the two variables. This means that, in general when tsteps is high, finscr is low, and vice versa.

	tsteps	finscr
tsteps	1.00000	
finscr	**-.74295**	1.00000

CRITICAL VALUE (1-TAIL, .05) = +/- .19833
CRITICAL VALUE (2-TAIL, .05) = +/- .23502

N = 70

Point Biserial Correlation Formula

The Point Biserial correlation is used when you are analyzing the relationship of a dichotomous variable, such as gender, and a continuous variable, such as the TAE Final Score. The dichotomous variable is categorized as either male or female and is assigned as either a 0 or 1 within the demographic database. The procedure for accomplishing this correlation can be found in most statistical handbooks.

MICROSTAT Job Aids

The next item of concern was to prepare job aids on the use of the MICROSTAT statistical package. The following examples were developed as job aids on how to use the program to:

1. Produce an analysis file from the performance database.

2. Compute descriptive statistics by use of the Descriptive Statistics function.

3. Compute Pearson correlations by producing a correlational matrix.

The example is to run some descriptive statistics on NAVMACS/SATCOM subsystem USH26 episode one. First we will enter the Data Management Subsystem's Recode/Transform/Select option to choose the variables that we want to output from STATFILE to a separate output file. After we have created the separate output file, then we can complete the descriptive statistics. Following the descriptive statistics example, is an example for completing a correlation matrix.

Creating a New Output File Within MICROSTAT

To begin, type

MICROSTAT

to enter the MICROSTAT package.Then you will see the Main MICROSTAT Menu.

MICROSTAT

OPTIONS:

A. DATA MANAGEMENT SUBSYSTEM	I. TIME SERIES ANALYSIS
B. DESCRIPTIVE STATISTICS	J. NONPARAMETRIC STATISTICS
C. FREQUENCY DISTRIBUTIONS	K. CROSSTAB / CHI-SQUARE TESTS
D. HYPOTHESIS TESTS: MEAN	L. PERMUTATIONS / COMBINATIONS
E. ANALYSIS OF VARIANCE	M. PROBABILITY DISTRIBUTIONS
F. SCATTERPLOT	N. HYPOTHESIS TESTS: PROPORTIONS
G. CORRELATION MATRIX	O. [Identification / Installation
H. REGRESSION ANALYSIS	P. Terminate]

ENTER: OPTION: __
From the Main Menu, select option A:

A. DATA MANAGEMENT SUBSYSTEM

to enter the Data Management Subsystem Menu, displayed below.

DATA MANAGEMENT SUBSYSTEM

DATA FILE OPTIONS:

A. ENTER DATA	H. DELETE CASES
B. LIST DATA	I. VERTICAL AUGMENT
C. EDIT DATA	J. SORT
D. RENAME FILE / EDIT HEADER	K. RANK-ORDER
E. FILE DIRECTORY	L. LAG TRANSFORMATIONS
F. DESTROY FILES	M. READ/WRITE EXTERNAL FILES
G. RECODE/TRANSFORM/SELECT	N. TRANSPOSE FILE
	O. [Terminate]

ENTER: OPTION: __

From the Data Management Subsystem menu, we want to select

G. RECODE/TRANSFORM/SELECT

to create a new output file for analyses within MICROSTAT.

Within the RECODE/TRANSFORM/SELECT option, you are immediately asked whether you have one or two input files. For our example, within TAE, let's select

A. ONE INPUT FILE

Give the name of our one main input file as

STATFILE

DATA MANAGEMENT SUBSYSTEM

G. RECODE/TRANSFORM/SELECT

OPTIONS: A. ONE INPUT FILE
B. TWO INPUT FILES

ENTER: OPTION: __

"STATFILE" is our student performance database within TAE. It contains performance information created when a subject takes a TAE episode and transformed through EDITVIEW. Each time we use the statistical data transformation file within EDITVIEW, a new "STATFILE.MSD" is created to be utilized within MICROSTAT analyses. The ".MSD" file extension indicates that this is a DATA FILE usable within the MICROSTAT package. Every time we use the data transformation capability within EDITVIEW, the TAE program writes over the previous version of "STATFILE.MSD."

Next, MICROSTAT will display all 673 variables present within STATFILE and then ask you about recoding/transforming your variables. We will answer "No" to this question.

601.	V601	602.	V602	603.	V603	604.	V604	605.	V605
606.	V606	607.	V607	608.	V608	609.	V609	610.	V610
611.	V611	612.	V612	613.	V613	614.	V614	615.	V615
616.	V616	617.	V617	618.	V618	619.	V619	620.	V620
621.	V621	622.	V622	623.	V623	624.	V624	625.	V625
626.	V626	627.	V627	628	V628	629.	V629	630.	V630
631.	V631	632.	V632	633.	V633	634.	V634	635.	V635
636.	V636	637.	V637	638.	V638	639.	V639	640.	V640
641.	V641	642.	V642	643.	V643	644.	V644	645.	V645
646.	V646	647.	V647	648.	V648	649.	V649	650.	V650
651.	V651	652.	V652	653.	V653	654.	V654	655.	V655
656.	V656	657.	V657	658.	V658	659.	V659	660.	V660
661.	V661	662.	V662	663.	V663	664.	V664	665.	V665
666.	V666	667.	V667	668.	V668	669.	V669	670.	V670
671.	V671	672.	V672	673.	V673				

WILL NEW VARIABLES BE CREATED WITH RECODE/TRANSFORMATIONS (Y,N)?

Then, MICROSTAT once again displays all of the 673 variables within STATFILE. We are prompted about the number of variables we want output. For our example, we will be outputting the 39 variables from STATFILE that make up the USH26 subsystem for episode one. Answer

39

to the variable output question.

r_{pb} = X

581.	V581	582.	V582	583.	V583	584.	V584	585.	V585
586.	V586	587.	V587	588.	V588	589.	V589	590.	V590
591.	V591	592.	V592	593.	V593	594.	V594	595.	V595
596.	V596	597.	V597	598.	V598	599.	V599	600.	V600
601.	V601	602.	V602	603.	V603	604.	V604	605.	V605
606.	V606	607.	V607	608.	V608	609.	V609	610.	V610
611.	V611	612.	V612	613.	V613	614.	V614	615.	V615
616.	V616	617.	V617	618.	V618	619.	V619	620.	V620
621.	V621	622.	V622	623.	V623	624.	V624	625.	V625
626.	V626	627.	V627	628.	V628	629.	V629	630.	V630
631.	V631	632.	V632	633.	V633	634.	V634	635.	V635
636.	V636	637.	V637	638.	V638	639.	V639	640.	V640
641.	V641	642.	V642	643.	V643	644.	V644	645.	V645
646.	V646	647.	V647	648.	V648	649.	V649	650.	V650
651.	V651	652.	V652	653.	V653	654.	V654	655.	V655
656.	V656	657.	V657	658.	V658	659.	V659	660.	V660
661.	V661	662.	V662	663.	V663	664.	V664	665.	V665
666.	V666	667.	V667	668.	V668	669.	V669	670.	V670
671.	V671	672.	V672	673.	V673				

ENTER: NUMBER OF VARIABLES TO BE OUTPUT (MAX= 923): 39
MICROSTAT then asks us about output specifications.

We'll select

A. OUTPUT ALL CASES (5)

OPTIONS: A. OUTPUT ALL CASES (5)

B. OUTPUT SUBSET OF CASES

C. SELECT INDIVIDUAL CASES

D. SELECT BY VALUE OF KEY VARIABLE

E. EXCLUDE BY VALUE OF KEY VARIABLE

ENTER: OPTION: _

Then enter the name of the new output file as

USH26E1

And, label the output file

USH26 EPISODE 1 VARIABLES

to better identify it.

OPEN FILE: C:STATFILE (PRESS RETURN TO USE OPEN FILE)

ENTER: NAME OF OUTPUT FILE: USH26E1

ENTER: FILE LABEL: USH26 EPISODE 1 VARIABLES
PRESS ANY KEY TO CONTINUE.

Next, enter the 39 variable numbers from STATFILE that you want included in the new USH26E1 output file. Enter variable number

2

through

40

MICROSTAT displays all the variables you enter so that you can ensure the output variables are correct.

24: 25. V25
25: 26. V26
26: 27. V27
27: 28. V28
28: 29. V29
29: 30. V30
30: 31. V31
31: 32. V32
32: 33. V33
33: 34. V34
34: 35. V35
35: 36. V36
36: 37. V37
37: 38. V38
38: 39. V39
39: 40. V40

OUTPUT SEQUENCE OK (Y,N)? YES

If you enter "N" indicating the output file is not correct, then you will enter all of the variable numbers again.

Once the variable numbers you input are fine, the data file USH26E1 will be output.

29: 30. V30
30: 31. V31
31: 32. V32
32: 33. V33
33: 34. V34
34: 35. V35
35: 36. V36
36: 37. V37
37: 38. V38
38: 39. V39
39: 40. V40

OUTPUT SEQUENCE OK (Y,N)? YES

FILE: C:USH26E1 IS NOW BEING OUTPUT. . .

DESTROY INPUT FILE C:STATFILE (N,Y)?

Answer

No

to the question about destroying the STATFILE input file.

Be careful when entering this response. Make sure that you always have a backup of your current working STATFILE.MSD just in case.

Executing Descriptive Statistics Within MICROSTAT

To begin, let's select

B. DESCRIPTIVE STATISTICS

from the Main MICROSTAT Menu.

OPTIONS:

A. DATA MANAGEMENT SUBSYSTEM	I. TIME SERIES ANALYSIS
B. DESCRIPTIVE STATISTICS	J. NONPARAMETRIC STATISTICS
C. FREQUENCY DISTRIBUTIONS	K. CROSSTAB / CHI-SQUARE TESTS
D. HYPOTHESIS TESTS: MEAN	L. PERMUTATIONS / COMBINATIONS
E. ANALYSIS OF VARIANCE	M. PROBABILITY DISTRIBUTIONS
F. SCATTERPLOT	N. HYPOTHESIS TESTS: PROPORTIONS
G. CORRELATION MATRIX	O. [Identification / Installation
H. REGRESSION ANALYSIS	P. [Terminate]

ENTER: OPTION: __

Then, we indicate the data file that we want to work with.

Enter

USH26E1

if it is not already open.

DESCRIPTIVE STATISTICS

OPEN FILE: C:USH26E1 (PRESS "RETURN" TO USE OPEN FILE)

ENTER: FILE NAME: _____

The 39 variables within the USH26E1 will then be displayed for you. Then, we want descriptive statistics on all cases, so select

A. INPUT ALL CASES

And select

A. SHORT FORM OUTPUT (MEAN, STD.DEV, MIN, MAX)

HEADER DATA FOR: C:USH26E1 LABEL: USH26 EPISODE 1 VARIABLES
ENTER: OPTION: _

NUMBER OF CASES: 5 NUMBER OF VARIABLES: 39

VARIABLE NUMBERS AND NAMES FOR: C:USH26E1

1.	V2	2.	V3	3.	V4	4.	V5	5.	V6
6.	V7	7.	V8	8.	V9	9.	V10	10.	V11
11.	V12	12.	V13	13.	V14	14.	V15	15.	V16
16.	V17	17.	V18	18.	V19	19.	V20	20.	V21
21.	V22	22.	V23	23.	V24	24.	V2	25.	V26
26.	V27	27.	V28	28.	V29	29.	V30	30.	V31
31.	V32	32.	V33	33.	V34	34.	V35	35.	V36
36.	V37	37.	V38	38.	V39	39.	V40		

OPTIONS: A. INPUT ALL CASES
B. INPUT SUBSET OF CASES

ENTER: OPTION: A

OPTIONS: A. SHORT FORM OUTPUT (MEAN, STD.DEV., MIN, MAX)
B. EXTENDED OUTPUT OF SELECTED VARIABLES
C. [Terminate]

Next, enter your output option and job title.

OPTIONS: A. SCREEN OUTPUT
B. PRINTER OUTPUT WITH FORMFEEDS
C. PRINTER OUTPUT WITHOUT FORMFEEDS
D. TEXT FILE OUTPUT

E. OUTPUT PRINTER SET-UP CODES
F. CHANGE PRINTER WIDTH. CURRENT VALUE: 80

ENTER: OPTION: B

ENTER: JOB TITLE:

USH26 EPISODE 1 VARIABLES._____

MICROSTAT will then display or print your descriptive statistics on the 39 USH26E1 variables (only 27 are shown here) depending on what option you have selected.

DESCRIPTIVE STATISTICS

HEADER DATA FOR: C:USH26E1 LABEL: USH26 EPISODE 1 VARIABLES
NUMBER OF CASES: 5 NUMBER OF VARIABLES: 39

USH26 EPISODE 1 VARIABLES.

NO.	NAME	N	MEAN	STD. DEV.	MINIMUM	MAXIMUM
1	V2	5	1.0000	.0000	1.0000	1.0000
2	V3	5	1.0000	.0000	1.0000	1.0000
3	V4	5	1.0000	.0000	1.0000	1.0000
4	V5	5	.0000	.0000	.0000	.0000
5	V6	5	.4000	.5477	.0000	1.0000
6	V7	5	.0000	.0000	.0000	.0000
7	V8	5	.0000	.0000	.0000	.0000
8	V9	5	.0000	.0000	.0000	.0000
9	V10	5	.0000	.0000	.0000	.0000
10	V11	5	1.0000	.0000	1.0000	1.0000
11	V12	5	.0000	.0000	.0000	.0000
12	V13	5	4.6000	2.0736	3.0000	8.0000
13	V14	5	.0000	.0000	.0000	.0000
14	V15	5	2.4000	2.6077	1.0000	7.0000
15	V16	5	.8000	.8367	.0000	2.0000
16	V17	5	.6000	.8944	.0000	2.0000
17	V18	5	.0000	.0000	.0000	.0000
18	V19	5	.0000	.0000	.0000	.0000
19	V20	5	2.0000	1.0000	1.0000	3.0000
20	V21	5	.0000	.0000	.0000	.0000
21	V22	5	.0000	.0000	.0000	.0000
22	V23	5	3.4000	2.4083	.0000	6.0000
23	V24	5	2.0000	1.2247	.0000	3.0000
24	V25	5	1.0000	.0000	1.0000	1.0000
25	V26	5	.0000	.0000	.0000	.0000
26	V27	5	.2000	.4472	.0000	1.0000
27	V28	5	.8000	.8367	.0000	2.0000

Once the data has been output you will see the Menu below displayed. From here you can continue using MICROSTAT for more computations or exit form the Descriptive Statistics function back to the Main Menu.

OPTIONS: A. REPEAT OUTPUT
B. MORE COMPUTATIONS
C. [Terminate]

ENTER: OPTION: __

Executing a Correlation Matrix Within MICROSTAT

Previously, we set up output files of important factor variables across episodes. We set up an output file with all the TAE test variables across episodes for Final Score and Incorrect LRUs to run analysis for Hypothesis 9. Hypothesis 9 states that "Subjects with a higher level of TS proficiency will have fewer incorrect solutions then less proficient subjects" (Conner, Hartley, & Mark, 1991).

To begin, select

G. CORRELATION MATRIX

from the Main MICROSTAT Menu.

MICROSTAT

OPTIONS:

A. DATA MANAGEMENT SUBSYSTEM	I. TIME SERIES ANALYSIS
B. DESCRIPTIVE STATISTICS	J. NONPARAMETRIC STATISTICS
C. FREQUENCY DISTRIBUTIONS	K. CROSSTAB / CHI-SQUARE TESTS
D. HYPOTHESIS TESTS: MEAN	L. PERMUTATIONS / COMBINATIONS
E. ANALYSIS OF VARIANCE	M. PROBABILITY DISTRIBUTIONS
F. SCATTERPLOT	N. HYPOTHESIS TESTS: PROPORTIONS
G. CORRELATION MATRIX	O. [Identification / Installation
H. REGRESSION ANALYSIS	P. [Terminate]

ENTER: OPTION: __
Enter the filename

HYPOTH 9

of the variables that you want correlated.

CORRELATION MATRIX

ENTER: FILE NAME: HYPOTH9

HEADER DATA FOR: C:HYPOTH9 LABEL: FINALSCR & INCLRU
NUMBER OF CASES: 5 NUMBER OF VARIABLES: 28

VARIABLE NUMBERS AND NAMES FOR: C:HYPOTH9

1. V40	2. V82	3. V124	4. V166	5. V208
6. V250	7. V292	8. V334	9. V376	10. V418
11. V460	12. V502	13. V544	14. V586	15. V27
16. V69	17. V111	18. V153	19. V195	20. V237
21. V279	22. V321	23. V363	24. V405	25. V447
26. V489	27. V531	28. V573		

OPTIONS: A. INPUT ALL CASES
B. INPUT SUBSET OF CASES

The variables within HYPOTH9 will then be displayed for you. We then indicate

A. INPUT ALL CASES

We want to correlate all the variables so we select

A. CORRELATE ALL VARIABLES

And title the correlation appropriately so it can be easily identified.

OPTIONS: A. CORRELATE ALL VARIABLES
B. CORRELATE SELECTED VARIABLES

ENTER: OPTION: a

ENTER: JOB TITLE:
HYPOTHESIS 9: CORRELATION--FINAL SCORE & INCORRECT LRUS._____

Next, we enter our display/print specifications.

OPTIONS: A. SCREEN OUTPUT
B. PRINTER OUTPUT WITH FORMFEEDS
C. PRINTER OUTPUT WITHOUT FORMFEEDS
D. TEXT FILE OUTPUT

E. OUTPUT PRINTER SET-UP CODES
F. CHANGE PRINTER WIDTH. CURRENT VALUE: 80

ENTER: OPTION: B

Then, select the correlation format that we want

A. OUTPUT CORRELATION MATRIX

OPTIONS: A. OUTPUT CORRELATION MATRIX
B. OUTPUT SSCP AND VAR-COVAR.
C. ALL OF THE ABOVE

ENTER: OPTION: _

Option A gives us a straight correlation matrix. Whereas, option B displays the raw sum of squares/cross products, adjusted sum of squares/cross products and variance/covariances in a tabular format. Option C will display both correlation formats from Option A and B.

The correlation matrix for HYPOTH9 is then displayed/printed for you by MICROSTAT. Only a part of the correlation matrix is shown below. Correlations greater than +/- .82 are significant at the .05 level of significance.

V195	-.37500	.87776	1.00000					
V237	.91856	-.25537	-.53583	1.00000				
V279	-.38851	.36062	.75275	-.55514	1.00000			
V321	-.30012	-.28377	.19294	-.34132	.78290	1.00000		
V363	-.37393	-.37424	.11218	-.42743	.72638	.98116	1.00000	
V405	.56728	-.70346	-.95406	.70530	-.87157	-.42453	-.37025	1.00000
V447	-.21539	-.59687	-.32308	.05862	.15342	.63411	.60136	.11848
V489	.87500	-.23205	-.25000	.73995	.02428	.19294	.11218	.33521
V531	.74984	-.22429	-.19849	.75630	.09425	.31393	.18472	.25477
V573	-.65661	.39969	.72501	-.60314	.81314	.61934	.54425	-.85220

	V447	V489	V531	V573
V447	1.00000			
V489	.03590	1.00000		
V531	.36101	.90422	1.00000	
V573	.39678	-.36934	-.09413	1.00000

CRITICAL VALUE (1-tail, .05) = +/- .82213
CRITICAL VALUE (2-tail, .05) = +/- .88233

N = 5

You can then continue with correlational computations or exit back to the Main Menu.

OPTIONS: A. REPEAT OUTPUT
B. MORE COMPUTATIONS
C. [Terminate]

ENTER: OPTION: __

CONCLUSIONS

The TAE system may have a number of uses within the training and fleet environments. One use relates to the long time period between graduation from ET "C" school and actual use of the knowledge and skills acquired in "C" school. TAE could be of importance in preventing the gradual deterioration of ETs' skills. The development of a variety of different fault scenarios across a wide range of equipment could be used to familiarize technicians with possible real-world situations. The scenarios could be developed to be ship-specific so that the information gained by the user would be of immediate significance. The possibility arises that, if an existing TAE scenario replicated a real-world situation, the scenario could be used as an aid in troubleshooting the real equipment. This would be especially useful if the equipment was difficult to reach or involved some risk to personnel.

Another use of TAE could be in the development of models of troubleshooting strategies. The behavioral profiles that are produced as a result of taking a TAE episode could be classified in terms of their usefulness in fault location. Once the profiles that result in fault location in an efficient manner are identified, they could be compared for strategies used to find the faulted LRU. From these comparisons, it might be possible to build a model of troubleshooting. It also might become apparent that different systems require different troubleshooting approaches. Results might show that different contextual factors such as combat conditions vs. peacetime conditions require different strategies.

These behavioral profiles also seem useful in identifying areas in which individuals need remedial training such as in the use of proof points to pinpoint faults. This seems especially useful to on-board supervisors who could use TAE to identify specific weaknesses in their personnel related to troubleshooting proficiency. Individuals could be given the TAE episodes and the results examined to identify weaknesses in approach to the fault-finding problem or in knowledge areas. Once these weaknesses are identified, the TAE could be used to demonstrate different troubleshooting strategies or to enlarge an individual's knowledge base. This possible use of TAE also demonstrates some of the advantages of TAE. It is a relatively simple system to operate and its ability to be used on a microcomputer makes it accessible anywhere there is an appropriate microcomputer. With the construction of an administrator's guide detailing instructions for use for a variety of conditions, the TAE system could be a powerful tool for both continuing and remedial training.

The TAE system could be used fleet-wide to assess the skills and knowledge that ETs are being equipped with in their various technical schools. Collection of fleet-wide data gained from the TAE episodes could be analyzed to discern skill and knowledge areas that show deficits across technicians. This could lead to changes in subject matter areas or to more emphasis on general problem solving strategies.

FUTURE EFFORTS

Performance Factor Definitions

The performance factor operational definitions should be consistent with the definitions expressed in the research design. There are actually two points in the data collection where the

performance factors are operationally defined. The first point is when the actual subject behaviors are recorded. The second point is when these behaviors are transformed into a format for data analysis. Care must be taken to ensure that, at both of these points in data collection, the operational definitions of the various factors correspond to the definitions as detailed in the supporting textual material.

Statistical Characteristics

The statistical characteristics of the performance factors will be investigated with intercorrelations between factors and factor analysis. The outcome of these analyses may show that some of the variance among TAE scores is accounted for by more than one factor. This will make it possible to reduce the set of performance factors for some statistical manipulations.

Since one method of assessing the reliability of the TAE system is by computing a correlation between the two scenarios within a subsystem, the scenarios should be reviewed to ensure that they are closely related. It might be necessary to choose different scenarios for testing to improve the reliability measure. There is a question about the extent to which the performance troubleshooting factors have been incorporated within the various episodes.

Composite Score

The individual factors need to be combined in a composite score. There are a number of different ways to combine factor scores into a composite score. One method might be more appropriate for obtaining a composite score, depending on the environmental context (i.e., peacetime vs. wartime). For example, if the composite score emphasized troubleshooting during wartime, total time and the supply limitation of replacing LRUs may affect the factor weightings in the composite score. Optional TAE factor composite scores should be explored.

Experimental Procedures

The experimental procedures need to be evaluated to ensure that no confounding variables have been introduced due to the experimental procedures. For example, due to the nature of troubleshooting proficiency experiment, all communication between the test administrator and subjects should be scripted and standardized. Any deviation from this standardized communication format could have a negative effect on the results.

It is extremely important that the TAE test administrator give explicit, directive instructions to the students regarding the extreme need for confidentiality regarding the answers to TAE episode tests. The debriefing could inform the students that the TAE test is not going to affect their school/ career performance and stress the need to keep all TAE testing activities confidential.

The TAE program does not appear to have a built-in mechanism for catching typing errors. It seems appropriate to emphasize to subjects that they proofread their inputs carefully. This caution applies especially to entering Reference Designator sites since a mistyped site could be recorded as an out-of-bounds response.

Data collection methods should be examined to ensure that the collected data is handled in a standardized format. There should be clearly stated reasons why experimental data are to be edited or ignored. One aid in determining if data are valid is for the test administrator to keep a daily log of unusual events or machine failures during test administration. Also, a log recording any data editions (within EDITVIEW) should be kept to verify the process.

DISTRIBUTION LIST

Distribution:
Director, Total Force Training and Education (OP-11)
Director, Training Technology (Code N-54)
Commanding Officer, Naval Education and Training Program Management Support Activity (Code 03)
Chief of Naval Technical Training (Code 00) (2)
Commander, Training Command, U.S. Atlantic Fleet
Commanding Officer, Service School Command, San Diego, CA
Defense Technical Information Center (2)

Copy to:
Commander, Naval Reserve Force, New Orleans, LA
Department of the Air Force, DET 5, Armstrong Laboratory Directorate, Brooks Air Force Base, TX
Commander, U.S. ARI, Behavioral and Social Sciences, Alexandria, VA (PERI-POT-I)
Superintendent, Naval Postgraduate School
Center for Naval Analyses