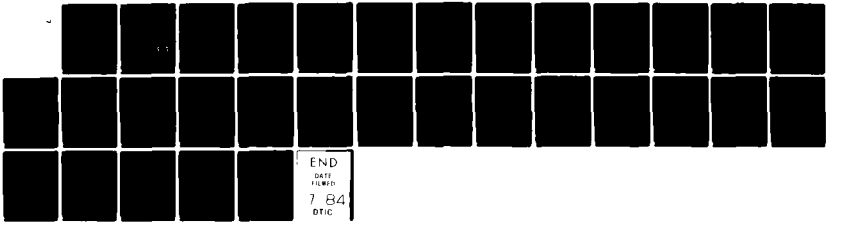
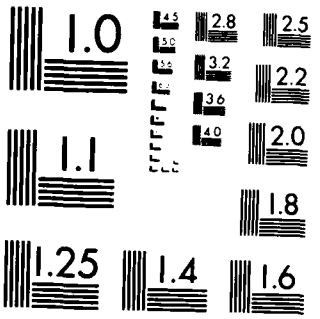
AD-A141 694 CONSTRUCTION OF A CRITERION-REFERENCED DIAGNOSTIC TEST FOR BOILER TECHNICIANS(U) NAVY PERSONNEL RESEARCH AND DEVELOPMENT CENTER SAN DIEGO CA G J LAABS ET AL. 1/0

UNCLASSIFIED MAY 78 NPRDC-TN-78-14 F/G 5/9 NL



END
DATE FILMED
7 84
DTIC



Technical Note 78-14 May 1978

(1)

CONSTRUCTION OF A CRITERION-REFERENCED, DIAGNOSTIC TEST FOR BOILER TECHNICIANS

Gerald J. Laabs
Robert C. Panell

Reviewed by
Adolph V. Anderson

Navy Personnel Research and Development Center
San Diego, California 92152

FOREWORD

This research and development was conducted in support of Exploratory Development Task Area ZF55.522.002 (Methodology for Development and Evaluation of Navy Training Programs). The criterion-referenced test developed and described in this report was used successfully in detecting deficiencies of Fleet personnel related to basic skills and knowledges essential to the operation and maintenance of the 1200 PSI Steam Propulsion Plant. The application of the test and findings are described in NPRDC TR 77-36, which is one of a series of six reports published in support of Advanced Development Subproject Z0108-PN.24, A Personnel Readiness Program.

Special appreciation is expressed to the Director, Propulsion Engineering School, Service School Command, Great Lakes, Illinois; to Mr. Hale Darling and Mr. Pete Tobarra of that command, for their help in test validation; and to BTC Harold T. Harris, Jr., BTC Danny L. Bowers, and MMC Jon Hall for their assistance in test development.

J. J. CLARKIN
Commanding Officer

SUMMARY

Problem

The development of a criterion-referenced, diagnostic test for Boiler Technicians (BTs) was a necessary part of a diagnostic testing/shipboard training program, the Personnel Readiness Training Program. In the absence of a well-defined technology for the construction of criterion-referenced tests, a systematic and practical approach to the development and evaluation of the diagnostic test was needed.

Purpose

The purpose of this effort was to develop a methodology for use in constructing the Basic Mechanical Procedures Test, a criterion-referenced, diagnostic test for Boiler Technicians.

Approach

Since the purpose of the test was to diagnose individual deficiencies, it was decided that it should be job related and keyed to the 14 BT training modules that had been adapted for shipboard use. The test was developed in two phases: an initial development phase followed by a refinement phase.

The initial phase included the following steps:

1. A pool of items based on known job requirements and maintenance documents typically encountered on the job was written by job experts.

2. The items were administered to pre- and postinstruction groups and those that best discriminated between the two groups were selected for each module.

3. A pass/fail criterion was determined for each module set, and used to classify students in a cross-validation sample. Validity for each set was then estimated by comparing actual group membership (pre- vs. postinstruction) to group membership assigned on the basis of diagnostic test scores.

The refinement phase included the following:

1. Additional items were constructed and original ones reformatted in accordance with a Navy Item Writing Manual.

2. Final test items were selected and validated using procedures outlined in 2 and 3 above.

3. Test-retest reliability was estimated using response data from two test administrations to BTs within the training/testing program.

Results

1. In the refinement phase, group classification agreements ranged from 68 to 92 percent, which showed that most of the module tests had excellent classification ability.

2. The discrimination ability of test items was improved by applying standards in the item writing manual.

3. For the two test administrations, group classification agreement was again very high, ranging from 71 to 96 percent. Classification agreements were statistically significant with the exception of one module, which approached significance.

Conclusions

1. High face and content validity was achieved by using materials that were encountered on the job and by having job experts write the items.

2. Test development and validity procedures not only resulted in a reliable and valid test, but are straightforward and easy to use.

3. The methodology developed provides a practical approach to the construction and evaluation of a criterion-referenced, diagnostic test.

CONTENTS

LIST OF TABLES

INTRODUCTION

Background

The Personnel Readiness Training Program, an Advanced Development effort, is concerned with the feasibility of using a diagnostic testing/shipboard training system to improve the readiness levels of Fleet personnel. In such a system, performance-oriented tests are used to diagnose deficiencies in job performance, and shipboard self-instructional materials are individually prescribed to correct deficiencies revealed by the diagnostic tests. Since the degree to which critical job skills can be improved through such a system may depend on the rating and/or the type of task involved, testing and training programs were developed for three applications: (1) the submarine Sonar Technician (ST) operating the AN/BQR-20A, (2) the submarine Missile Technician (MT) operating the Missile Test and Readiness Equipment (MTRE Mk 7 Mod 2), and (3) the Boiler Technician (BT) operating and maintaining the 1200 PSI Steam Propulsion Plant.

A series of five reports have been published concerning the Personnel Readiness Training Program. The first in the series--Laabs, Main, Abrams, & Steinemann, Note 1--described the general approach of the program and how it was being applied in the three areas. The next three--Winchell, Panell, and Pickering, 1976; Laabs, Panell, and Pickering, 1977; and Laabs, Harris, and Pickering, 1977--provided descriptions of the ST, MT, and BT applications. The final report--Anderson, Laabs, Pickering, and Winchell, 1977--summarized findings and conclusions across applications.

Problem

The BT application required the development of the Basic Mechanical Procedures Test, a criterion-referenced, diagnostic test that is keyed to individualized, self-paced instruction. A criterion-referenced test differs from the norm-referenced test, which is traditionally used within the standard instruction model, in several ways. First, it compares a student's performance against a standard or criterion, instead of against the performance of another student. Thus, it is appropriate for use in an individualized instruction program, which usually involves the assessment of a student's "absolute" level of skill or knowledge for such purposes as diagnosing instructional needs or deciding which sequence of information should be followed. A norm-referenced test does not yield this type of information.

Another distinction between the two types of tests is based upon the way test scores are used or interpreted (Hambleton & Novick, 1973). For example, in a norm-referenced application, the results of, say, a typing test would be used to determine the standing of one individual in relation to others. In a criterion-referenced application, the results of the same test would be used to decide whether or not more practice or training is needed.

Finally, the two types of tests differ in test development procedures. For example, for a norm-referenced test, the items selected are generally those that only some of the students can be expected to answer correctly after they have completed instruction. Standard measures of reliability and validity for tests composed of such items can then be based upon the

variability in final test scores obtained. Conversely, for a criterion-referenced test, the items selected are those that all of the students can be expected to answer correctly after, but not before, they have completed instruction. Thus, the standard measures of reliability and validity are not applicable for tests composed of these items, because the variability in final test scores obtained is constrained (Popham & Husek, 1969).

For these reasons, the technology for the construction and evaluation of norm-referenced tests is not applicable for criterion-referenced tests.

Purpose

The purpose of this effort was to develop a methodology for constructing and evaluating a criterion-referenced, diagnostic test that is keyed to individualized, self-paced instruction.

APPROACH

The Propulsion Engineering School, Service School Command, Great Lakes recently adopted a new curriculum for BTs, consisting of modularized, self-paced instruction, based on a thorough task analysis of the BT's job (Brock & DeLong, 1975). Fourteen of these modules, which are listed in Table 1, were adapted for shipboard use by the Naval Education and Training Support Center, Pacific, San Diego.

Table 1

Basic Skills and Knowledges Modules

Module	Title
1	Metal Fasteners, Hand Tools
2	Pipes, Tubing, Fittings
3	Packing, Gaskets, Insulation
4	Valves
5	Bearings, Lubrication
6	Pumps
7	Precision Measuring Instruments, Technical Manuals
8	Heat Properties, Heat Exchangers
9	Indicating Devices
10	Turbines, Couplings, Gears
11	Strainers, Purifiers
12	Low Pressure Air System and Compressor
13	Oil Pollution
14	Planned Maintenance System

The purpose of the criterion-referenced Basic Mechanical Procedures Test was to diagnose individual deficiencies of BTs so that remedial training could be assigned; thus, it was decided that the test should be job-related and keyed to the 14 training modules adapted for shipboard use. Since the adapted modules were well designed, they were accepted without change for use in this effort.

The final test was to require no more than 1-1/2 hours to complete. It was developed in two phases: an initial development phase followed by a refinement phase. These phases are described in the following section.

TEST DEVELOPMENT

Initial Development Phase

Item Construction

A pool of 186 items (approximately twice the number needed) was constructed by (1) setting up hypothetical job situations that required the supporting skills and knowledges covered in one or more of the modules, (2) relating a given situation to supporting skills and knowledges, and (3) writing questions about each situation. A team of job experts (i.e., three Chief Petty Officers) assisted in this procedure.

To ensure that the test would be job-related, the hypothetical job situations were based on known job requirements obtained primarily from information in Maintenance Requirement Cards (MRCs), which document periodic maintenance tasks to be performed. Additional information was obtained from a BT task analysis (see Brock & DeLong, 1975) and the Personnel Qualifications Standards (PQS), which is a training document. Each job situation set up referred to one of the following:

1. MRCs, an example of which is provided in Figure 1.

2. Charts and diagrams supporting a particular maintenance action or job.

3. Illustrations of equipments or parts of equipments, without specifying a particular maintenance action or job.

An example of a situation set up by referring to the MRC shown in Figure 1 and a question corresponding to that situation is shown below:

> SITUATION: You are assigned the Maintenance Requirement Card (MRC) shown on the opposite page, which lists routine maintenance procedures to be performed on the turbine of a fuel oil service pump.
>
> QUESTION: The purpose of the strap wrench listed on the MRC is to:
>
> a. Remove the plug from the thrust bearing cover plate.
> b. "Turn over" the pump shaft without damaging its surface.
> c. Prevent damage to nuts, bolts, and other fittings.
> d. Remove the straps from the pump couplings.

Information needed to answer this question correctly is contained in one of the frames of the self-instructional material. Other questions asked under this situation tested the examinee's ability to read and interpret the card.

SYSTEM	COMPONENT	MRC CODE
Propulsion	Fuel Oil Service Pump	F-4 Q-5

SUBSYSTEM	RELATED MAINTENANCE	RATES	M-H
Fuel Oil Service	None	BT2 FN	0.3 0.3

MAINTENANCE REQUIREMENT DESCRIPTION

1. Measure turbine thrust clearance.

TOTAL M-H 0.6
ELAPSED TIME 0.3

SAFETY PRECAUTIONS

1. Observe standard safety precautions.
2. Wire steam inlet and exhaust valves shut and tag "Do Not Open."

TOOLS, PARTS, MATERIALS TEST EQUIPMENT

1. Safety tags
2. Strap wrench
3. 24 Gauge wire
4. Paint scraper
5. Extension light
6. Pencil and paper
7. 12" Adjustable wrench
8. 0"-1" Depth micrometer

PROCEDURE

Preliminary

a. Wire steam inlet and exhaust valves shut and tag "Do Not Open".
b. Ensure turbine has been idle at least 24 hours.

1. Measure Turbine Thrust Clearance.
   a. Remove paint from thrust bearing cover plate.
   b. Remove plug from thrust bearing cover plate.
   c. Attach strap wrench to pump coupling; rotate pump shaft 1/4 turn clockwise.
   d. Measure distance from thrust bearing cover plate to thrust bearing locknut; record readings.
   e. Rotate pump shaft 1/4 turn counterclockwise.
   f. Measure distance from thrust bearing cover plate to thrust bearing locknut; proper turbine thrust clearance is minimum 0.010", maximum 0.025".
   g. Reinstall plug in thrust bearing cover plate.
   h. Remove strap wrench.
   i. Remove wire and safety tags from steam inlet and exhaust valves.

PAGE 1 OF 1

60 X357 Q

LOCATION	DATE
	June 1970

Figure 1. Example of a Maintenance Requirement Card (MRC).

Item Administration and Selection

The 186 items were administered to two groups of BTs assigned to the shore-based Propulsion Engineering School. The first group consisted of 100 BTs who were entering the course (Preinstruction Group) and the second, of 100 BTs who had completed the course (Postinstruction Group). Response data for 25 students randomly selected from each group were put aside for use in cross-validation. Response data for the remaining 75 students in each group were used to perform an item analysis.

In the traditional item analyses used for norm-referenced tests, the differences among individuals are maximized by selecting items that were answered correctly by about half of the respondees, and/or had responses that correlated highly with the total test score. Using these procedures, items that most individuals answer correctly or incorrectly are discarded because they add nothing to the test score variability, which is the basis for estimating test reliability and validity. Criterion-referenced tests, on the other hand, are designed to determine if individuals possess a certain skill or knowledge rather than to discriminate among those at the same general skill or knowledge level. Thus, for the BT diagnostic test, items were required that discriminated between those who needed to study the modules and those who did not, rather than items that maximized differences among those who had the information contained in the modules. The most direct way of selecting such test items was to select those that showed the largest difference in difficulty between the Preinstruction and Postinstruction Groups; that is, they were maximally sensitive to instructional gain. Thus, the items selected for the Basic Mechanical Procedures Test were those that were answered correctly by (1) a significantly greater proportion of Postinstruction Group members than Preinstruction Group members (i.e., $p < .05$ for a t-test of proportions), and (2) at least 50 percent of the Postinstruction Group. Of the total of 186 items, 101 met these criteria.

Determination of Pass/Fail Criterion

An effort was made to minimize the error of assigning training to BTs who did not need it, since this would not only damage the credibility of the testing/training program but also waste valuable training time. Thus, the pass/fail criterion for a set of questions selected for a given module was determined by:

1. Calculating (a) the proportion of correct responses for each question in the set given by Postinstruction Group members, and (b) the average proportion of correct responses for the entire question set given by these members.

2. Calculating a .95 confidence interval about the average proportion correct (b above).

3. Multiplying the lower cutoff of the confidence interval by the number of questions in the set, and using the nearest whole number as the pass/fail criterion. The lower cutoff of the confidence interval was used to ensure that modules would be assigned only to those who absolutely needed

the training (even if it meant that modules would not be assigned to those who might be able to use them).

For example, a set of 10 questions was selected for Module 1. The average proportion of correct responses for Postinstruction Group members on this set was .82, and the lower cutoff of the .95 confidence interval about this proportion was .74. Thus, the pass/fail criterion was determined as follows: 10 x .74 = 7.4 or 7. An individual who had seven or more questions in this set correct would be classified as not needing instruction; and one who had less than seven correct, as needing instruction.

Test Validity

Very little work has been done on establishing the validity of criterion-referenced tests; usually only face and content validity have been considered. The Basic Mechanical Procedures Test has both high face and content validity for the following reasons:

1. The test questions were directly related to BT job situations.

2. Job experts assisted in the test development process by determining what information contained in the modules was needed for a given job and by writing test questions.

3. Extensive use was made of MRCs, charts, diagrams, and illustrations in presenting each job situation and its associated questions.

A further measure of validity was obtained by using the pass/fail criterion for each module (with the exception of Module 8, which had no acceptable questions) to classify the students in the cross-validation sample, as would be done when using the test as a diagnostic instrument. If a student's performance was at or above the criterion, he was classified as a Postinstruction Group member; if his performance was below the criterion, he was classified as a Preinstruction Group member. This classification was then compared with actual group membership, and the percent agreement between actual group membership and membership assigned on the basis of test scores obtained was determined. Results are provided in Table 2, which shows that, overall, the module tests had very good to excellent classification ability. The percent agreements ranged from 59 percent for Module 5 to 92 percent for Modules 6 and 7; all agreements except that for Module 5 were significant at conventional levels, as shown by a Chi Square test with Yates correction for continuity. Appendix A contains the contingency tables and cross-validation data for each of the modules.

Table 2

Agreement in Classification of a Cross-Validation Sample (Phase I)
(N = 49)

Module	Title	Number Items Selected	Pass/Fail Criterion	Percent Agreement
1	Metal Fasteners, Hand Tools	10	7	76
2	Pipes, Tubing, Fittings	4	3	75
3	Packing, Gaskets, Insulation	5	3	67
4	Valves	20	13	76
5	Bearings, Lubrication	5	3	59
6	Pumps	13	9	92
7	Precision Measuring Instruments, Technical Manuals	10	7	92
8	Heat Properties, Heat Exchangers	0	--	--
9	Indicating Devices	4	2	73
10	Turbines, Couplings, Gears	7	5	82
11	Strainers (Lesson 1)[a]	6	4	84
12	Low Pressure Air System and Compressor	8	6	86
13	Oil Pollution	4	3	78
14	Planned Maintenance System	5	4	84
		101		

Note. The final cross-validation sample consisted of 49 students since response data for one Postinstruction Group member were not useable.

[a]Questions regarding purifiers were omitted because they pertained to machinist's mates only.

Refinement Phase

Additional Item Construction/Modification

To provide an adequate set of questions for Modules 5 and 8 and to strengthen the other sets, 31 new questions were constructed in the same manner as before. However, this time, the format rules contained in an item writing manual (NAVEDTRAPRODEVCENINST 1552.1), which was promulgated after the original items were constructed and validated, were followed. Also, to further improve the test, the 96 items found acceptable after cross-validation[1] were compared to format standards contained in the manual, and 32 violations were discovered. Most of these item writing violations pertained to the following four main format rules:

1. Capitalize a negative word in the stem.
2. Place qualifying information in the beginning of the stem.
3. Word the stem so that there is no doubt about what is being asked.
4. Place as much of the wording as possible into the stem.

As a result, the 32 items containing these violations were slightly reworded. Changes were made in format only; in no case was the content of these items or the alternative answers changed. The effect of these format changes is discussed below.

Item Administration and Selection

The 127 items were administered to two groups of BTs assigned to the shore-based school. This time, the Preinstruction Group consisted of 75 BTs who were entering the course; and the Postinstruction Group, of 75 BTs who had completed it. Again, response data from 25 students randomly selected from each group was put aside for cross-validation purposes. The data from the remaining 50 students in each group was used to conduct an item analysis. The method used was the same as that used in the initial development phase. As a result, 85 of the original 96 items were retained and 23 of the 31 new items were selected. Thus, the final test consisted of 108 items.

Effect of Format Changes

Previous efforts that have examined the effect of format changes (e.g., Board & Whitner, 1972; Dunn & Goldstein, 1959; McMorris, Brown, Snyder, & Pruzek, 1972) have been devoted primarily to the effect of "incorrectly" written items on norm-referenced test statistics (e.g., item difficulty, validity coefficients, and reliability coefficients). These statistics, of course, do not apply to the criterion-referenced Basic Mechanical Procedures Test, which was developed to discriminate between BTs who do and do not require information contained in the 14 modules. Therefore, the effect of format changes to 32 of the original items was examined by analyzing their discrimination ability.

---

[1]The five items previously selected for Module 5 were discarded.

Because of the small number of violations of each of the four main format rules listed above, an overall analysis was performed. This was done by comparing the overall scores obtained by the Preinstruction and Postinstruction Groups on the 32 original items during the initial development phase with those obtained by the two similar groups on the 32 reformatted items during this (refinement) phase. The difference was then submitted to a Wilcoxon Matched-Pairs Signed-Ranks Test, which showed a significant effect, $z = 2.73$, $p < .01$. Thus, it is clear that application of the item writing guide did improve the test. However, some of the items showed very small or even negative difference scores.

Determination of Pass/Fail Criterion

The method used to determine the pass/fail criterion for each set of questions was the same as that used in the initial test development phase.

Test Validity

The final Basic Mechanical Procedures Test had high face and content validity, for the reasons listed previously.

As in the initial development phase, a further measure of validity was obtained by using the new pass/fail criterion calculated for each module to classify the students in the cross-validation sample. This classification was compared to actual group membership, and the percentage of agreement was determined. Results are provided in Table 3, which shows that most of the module tests showed excellent classification ability. The percent agreements ranged from 68 percent for Module 3 to 92 percent for Module 6; all were significant at conventional levels, as determined by a Chi Square test with Yates correction for continuity. Appendix B contains the contingency tables and cross-validation data for each of the modules.

Test-Retest Reliability

In a program designed to diagnose deficiencies and to assign training to remedy those deficiencies, test-retest reliability can be viewed in terms of the consistency of the diagnostic decisions made at two different points in time. Thus, within the BT testing/training program, the final version of the Basic Mechanical Procedures Test was administered to a group of 28 BTs. These BTs received no diagnostic feedback or remedial training, and were administered the test a second time from 3 to 6 months later. This procedure provided response data for use in estimating test-retest reliability.

Table 3

Agreement in Classification of a Cross-Validation Sample (Phase II)
(N = 50)

Module	Title	Number Items Selected	Pass/Fail Criterion	Percent Agreement
1	Metal Fasteners, Hand Tools	8	7	88
2	Pipes, Tubing, Fittings	6	3	78
3	Packing, Gaskets, Insulation	5	3	68
4	Valves	17	11	86
5	Bearings, Lubrication	8	4	74
6	Pumps	12	9	92
7	Precision Measuring Instruments, Technical Manuals	10	7	86
8	Heat Properties, Heat Exchangers	4	2	72
9	Indicating Devices	7	5	78
10	Turbines, Couplings, Gears	7	5	80
11	Strainers (Lesson 1)	6	4	88
12	Low Pressure Air System and Compressor	8	6	82
13	Oil Pollution	4	3	86
14	Planned Maintenance System	6	4	90
		108		

Although the proportion of examinees consistently classified in the same category across two test administrations has been suggested as a measure of such reliability, this procedure does not account for the agreement expected by chance alone (Swaminathan, Hambleton, & Algina, 1974). Therefore, to determine the percent agreement in diagnostic decisions over the two administrations of the Basic Mechanical Procedures Test, the coefficient Kappa (K) was used. This coefficient, which was introduced by Cohen (1960), does consider chance agreement and is defined as

$$K = (P_{ob} - P_c)/(1 - P_c)$$

where $P_c = P_o^2$ and $P_{ob} = P_{oo} + P_{11}$

are taken from the matrix shown below:

		PRETEST PASS	PRETEST FAIL	
POSTTEST	PASS	P_{oo}		P_o
	FAIL		P_{11}	P_1

A quality of this coefficient is that it indexes the degree of agreement rather than the degree of association, as is determined by Phi or Chi-square. Thus, the Phi and Chi-square values increase with discrepancies between observed and chance or expected values, regardless of whether these discrepancies are in the direction of agreement or disagreement. Since the Kappa statistic does not consider the degree of association in the disagreement cells, it provides a better measure of agreement (Cohen, 1968).

There are no standardized tables of the significance of Kappa. Therefore, for purposes of this study, the level of significance was determined by dividing K by σ_{k_o}, with σ_{k_o} defined as

$$\sigma_{k_o} = \sqrt{\frac{P_c}{N(1-P_c)}}$$

and referring the resulting $\underline{z}$ value to the normal curve tables. Results are presented in Table 4, which indicates that the test was highly reliable, considering the length of time between the two administrations. All of the modules reached conventional levels of significance with the exception of Module 14, which approached significance and showed 75 percent agreement.

Table 4

Reliability of Diagnostic Decisions for Two Test Administrations
(N = 28)

Module	Title	Percent Agreement	Kappa Coefficient	p
1	Metal Fasteners, Hand Tools	89	.70	.001
2	Pipes, Tubing, Fittings	86	.59	.015
3	Packing, Gaskets, Insulation	79	.52	.01
4	Valves	86	.58	.015
5	Bearings, Lubrication	71	.38	.03
6	Pumps	82	.44	.05
7	Precision Measuring Instruments, Technical Manuals	75	.49	.01
8	Heat Properties, Heat Exchangers	71	.36	.05
9	Indicating Devices	71	.38	.03
10	Turbines, Couplings, Gears	71	.43	.012
11	Strainers (Lesson 1)	96	.89	.001
12	Low Pressure Air System and Compressor	82	.50	.025
13	Oil Pollution	71	.39	.025
14	Planned Maintenance System	75	.30	.12

DISCUSSION AND CONCLUSIONS

A systematic approach was developed and followed in the construction of a criterion-referenced, diagnostic test keyed to an individualized, self-paced instruction program for Boiler Technicians. The main steps involved in the approach were: (1) writing an item pool, (2) selecting test items, (3) determining cutoff scores, (4) estimating validity, and (5) estimating test-retest reliability. Use of this approach resulted in the development of a reliable and valid test.

Several factors account for the high reliability and validity of the Basic Mechanical Procedures Test. First, as indicated previously, it had high face and content validity for the following reasons:

1. Job requirements were used to set up hypothetical job situations.

2. Illustrations and reproductions of cards, charts, and diagrams that would be encountered on the job were used to present the situations.

3. Job experts assisted in the development process by determining what information contained in the instruction program would be needed to perform the job described and by writing questions about each situation.

Second, strict rules were followed in selecting items for inclusion on the test and in determining cutoff scores. Only those items that showed significant instructional gain between preinstruction and postinstruction were chosen. Cutoff scores for separate parts of the test were determined on the basis of the performance of postinstruction groups and were calculated to minimize the error of assigning training to a BT who did not need it.

Third, further estimates of validity were made using cross-validation samples of preinstruction and postinstruction groups. These estimates were calculated by comparing actual group membership to membership assigned on the basis of the diagnostic test scores on each part of the test. Reliabilities were estimated by the proportion of examinees classified in the same category on each part of the test across two diagnostic test administrations. Finally, the Kappa statistic was used to assess the statistical significance of the classification agreements. These test development and evaluation procedures not only result in a reliable and valid test, but also have the advantage of being straightforward and easy to use.

Overall, the methodology outlined above provides a practical approach to the construction of a criterion-referenced, diagnostic test. Its use, however, is limited to the situation where the instructional material to which the test is being keyed is known to be effective. In the case where instructional material is being evaluated and could be changed or revised, the item selection procedures would be based on the course objectives. That is, an item that does not show large instructional gain between pre-instruction and postinstruction would not be discarded or rewritten until the possibility is eliminated that the course did not adequately cover the objective being tested by the item.

REFERENCES

Anderson, A. V., Laabs, G. J., Pickering, E. J., & Winchell, J. D. A personnel readiness training program: Final report (NPRDC Tech. Rep. 77-39). San Diego: Navy Personnel Research and Development Center, August 1977. (AD-A043 371)

Board, C., & Whitner, D. R. The effect of selecting poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 1972, 9, 225-233.

Brock, J. F., & DeLong, J. L. Design and conduct of a mechanical maintenance training program with an annual flow rate of 11,000 trainees: A review. In W. A. King and J. Duva (Eds.), New concepts in maintenance training and performance aids (NTEC TH-255). Orlando, FL: Human Factors Laboratory, Naval Training Equipment Center, 1975.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin, 1968, 70, 213-220.

Dunn, T. F., & Goldstein, L. G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. Educational and Psychological Measurement, 1959, 19, 171-179.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Laabs, G. J., Harris, H. T., Jr., & Pickering, E. J. A personnel readiness training program: Operation and maintenance of the 1200 PSI Steam Propulsion Plant (NPRDC Tech. Rep. 77-36). San Diego: Navy Personnel Research and Development Center, June 1977. (AD-A042 033)

Laabs, G. J., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Maintenance of the missile test and readiness equipment (MTRE Mk 7 Mod 2) (NPRDC Tech. Rep. 77-19). San Diego: Navy Personnel Research and Development Center, March 1977. (AD-A037 546)

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

McMorris, R. F., Brown, J. A., Synder, G. W., & Pruzek, R. M. Effects of violating item construction principle. Journal of Educational Measurement, 1972, 9, 287-295.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Winchell, J. E., Panell, R. C., & Pickering, E. J. A personnel readiness training program: Operation of the AN/BQR-20A (NPRDC Tech. Rep. 77-4). San Diego: Navy Personnel Research and Development Center, November 1976. (AD-A033 435)

REFERENCE NOTE

1. Laabs, G. J., Main, R. E., Abrams, A. J., & Steinemann, J. H. A personnel readiness training program: Initial project developments (NPRDC Special Rep. 75-8). San Diego: Navy Personnel Research and Development Center, April 1975.

APPENDIX A

FIRST VALIDATION: CONTINGENCY TABLES AND CROSS-VALIDATION DATA

FIRST VALIDATION: CONTINGENCY TABLES AND CROSS-VALIDATION DATA

Module 1 Criterion = 7/10 % Agreement = 76 $\underline{X}^2 = 11.16$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	17	8
Post	4	20

Module 2 Criterion = 3/4 % Agreement = 75 $\underline{X}^2 = 10.33$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	19	5
Post	7	19

Module 3 Criterion = 3/5 % Agreement = 67 $\underline{X}^2 = 4.58$, $\underline{p} < .05$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	17	8
Post	8	16

Module 4 Criterion = 13/20 % Agreement = 76 $\underline{X}^2 = 11.91$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	23	2
Post	10	14

Module 5 Criterion = 3/5 % Agreement = 59 $\underline{X}^2 = .99$ NS

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	18	7
Post	13	11

Module 6 Criterion = 9/13 % Agreement = 92 $\underline{X}^2 = 31.22$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	24	1
Post	3	21

Module 7 Criterion = 7/10 % Agreement = 92 $\underline{X}^2 = 31.27$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	22	3
Post	1	23

Module 8 No Acceptable Questions

Module 9 Criterion = 2/4 % Agreement = 73 $\underline{X}^2 = 11.50$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	13	12
Post	1	23

Module 10 Criterion = 5/7 % Agreement = 82 $\underline{X}^2 = 17.47$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	19	6
Post	3	21

Module 11 Criterion = 4/6 % Agreement = 84 $\underline{X}^2 = 19.20$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	20	5
Post	4	20

Module 12 Criterion = 6/8 % Agreement = 86 $\underline{X}^2 = 22.27$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	3	21

Module 13 Criterion = 3/4 % Agreement = 78 $\underline{X}^2 = 13.00$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	18	7
Post	4	20

Module 14 Criterion = 4/5 % Agreement = 84 $\underline{X}^2 = 19.77$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	20	5
Post	3	21

APPENDIX B

SECOND VALIDATION: CONTINGENCY TABLES AND CROSS-VALIDATION DATA

SECOND VALIDATION: CONTINGENCY TABLES AND CROSS-VALIDATION DATA

Module 1 Criterion 7/8 % Agreement = 88 $X^2 = 26.09$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	2	23

Module 2 Criterion = 3/6 % Agreement = 78 $X^2 = 16.77$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	14	11
Post	0	25

Module 3 Criterion = 3/5 % Agreement = 68 $X^2 = 5.12$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	17	8
Post	8	17

Module 4 Criterion = 11/17 % Agreement = 86 $X^2 = 23.16$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	3	22

Module 5 Criterion = 4/8 % Agreement = 74 $X^2 = 9.69$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	19	6
Post	7	18

Module 6 Criterion = 9/12 % Agreement = 92 $X^2 = 32.21$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	24	1
Post	3	22

Module 7 Criterion = 7/10 % Agreement = 86 $X^2 = 23.16$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	3	22

Module 8 Criterion = 2/4 % Agreement = 72 $X^2 = 8.21$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	16	9
Post	5	20

Module 9 Criterion = 5/7 % Agreement = 78 $X^2 = 13.54$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	19	6
Post	5	20

Module 10 Criterion = 5/7 % Agreement = 80 $X^2 = 16.64$, $p < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	17	8
Post	2	23

Module 11 Criterion = 4/6 % Agreement = 88 $\underline{X}^2 = 26.60$, $\underline{p} < .01$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	20	5
Post	1	24

Module 12 Criterion = 6/8 % Agreement = 88 $\underline{X}^2 = 26.09$, $\underline{p} < .01$

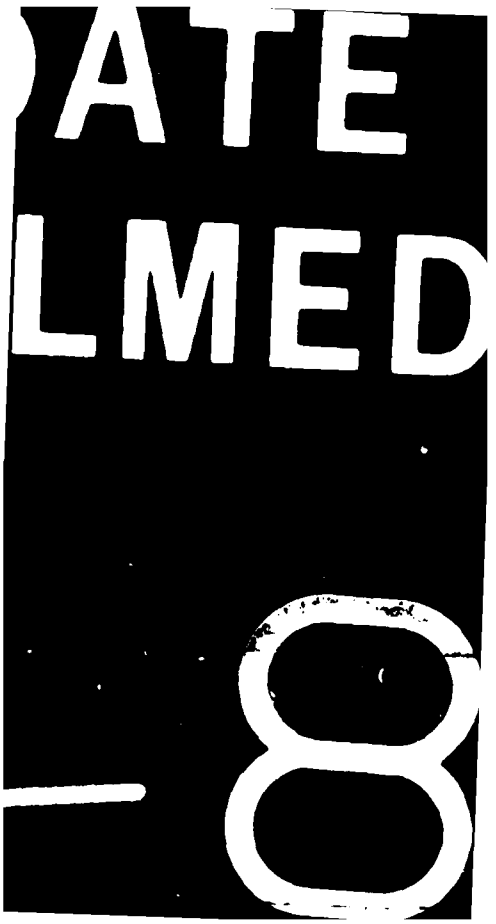
Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	2	23

Module 13 Criterion = 3/4 % Agreement = 86 $\underline{X}^2 = 24.08$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	19	6
Post	1	24

Module 14 Criterion = 4/6 % Agreement = 90 $\underline{X}^2 = 29.30$

Diagnosed Group Membership	Actual Group Membership: Pre	Actual Group Membership: Post
Pre	21	4
Post	1	24


ATE
LMED
8