DOCUMENT RESUME

ED 227 356 CE 035 459

AUTHOR Claudy, John G.; Appleby, Judith A.
TITLE Validating Competency Tests and Using Test Results. Module 20. [Vocational Education Curriculum Specialist.]
INSTITUTION American Institutes for Research in the Behavioral Sciences, Palo Alto, Calif.
SPONS AGENCY Office of Vocational and Adult Education (ED), Washington, DC.
PUB DATE 82
CONTRACT 300-79-0735
NOTE 39p.; For related documents, see ED 215 114-132 and CE 035 456-458.
AVAILABLE FROM East Central Network, Sangamon State University, E-22, Springfield, IL 62708 ($4.00; complete set--20 modules, instructor's guide, audio cassette, field test report--$45.00).
PUB TYPE Guides - Classroom Use - Materials (For Learner) (051)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Behavioral Objectives; Competence; Competency Based Education; Inservice Education; Job Skills; Learning Activities; Learning Modules; Postsecondary Education; Predictive Validity; Secondary Education; *Standards; Student Evaluation; Test Results; *Test Use; *Test Validity; *Vocational Education; Vocational Education Teachers
IDENTIFIERS *Competency Tests; *Curriculum Specialists

ABSTRACT

This module, the fourth of four units about vocational competency measurement, is module 20 in the Vocational Education Curriculum Specialist series. The purpose stated for the document is to help in the validation of a competency test after it has been developed, in the determination of how test results will be reported, and in the consideration of ways of setting standards for passing or failing the test. Content is organized into three sections, each of which focuses on one goal and two or more objectives. The first section discusses possible approaches for determining the content validity of a competency test. It also comments on predictive validity and maintaining test validity. Section 2 considers possible uses for vocational competency tests and how test results can be reported in accordance with the intended use. The module concludes with the question of how to set standards for passing or failing a test. Each section concludes with individual study activities, discussion questions, and group activities. Self-check items and possible responses to them are appended for use as a pretest and review of the module content. (YLB)

VALIDATING COMPETENCY TESTS AND USING TEST RESULTS

Module 20

John G. Claudy

with the assistance of Judith A. Appleby

Developed by the American Institutes for Research under support from the Office of Vocational and Adult Education, U.S. Department of Education. 1982.

The VECS materials are printed and distributed by
East Central Network for Curriculum Coordination
Sangamon State University, E-22
Springfield, IL 62708
217/786-6375

The information reported herein was obtained pursuant to Contract No. 300-79-0735 with the U.S. Department of Education. Contractors undertaking such projects under government sponsorship are encouraged to document information according to their observation and professional judgment. Consequently, information, points of view, or opinions stated do not necessarily represent official Department of Education position or policy.

Table of Contents

ACKNOWLEDGMENTS

The discussions and techniques presented in Modules 17 through 20 of the VECS series are based on the work of the American Institutes for Research in carrying out the Vocational Competency Measures (VCM) project under contract with the Office of Vocational and Adult Education, U. S. Department of Education. The project was a major effort, beginning in October 1979 and continuing through 1982, to provide a national model for vocational competency test development.

The VCM project had four major objectives:

(1) To develop competency tests in selected occupations, representing each of the seven major areas: trade and industry, home economics, health, distributive education, technical, business and office, and agriculture;

(2) To establish their usefulness through extensive field testing and evaluation;

(3) To promote their acceptance and use in vocational education programs;

(4) To design and help implement a program for continuing occupational competency test development on a self-supporting basis.

The successful implementation of the project was due to the efforts of many people. Senior project staff responsible for specific tasks were:

Dr. Albert B. Chalupsky, Project Director
Ms. Marion F. Shaycoft, Director of Sampling and Test Quality Control, and Test Team Leader
Dr. Malcolm N. Danoff, Director of Field Coordination and Validation
Dr. Robert A. Weisgerber, Director of Competency Requirements Analysis and Test Team Leader
Ms. Judith A. Appleby, Director of Dissemination
Dr. John G. Claudy, Test Team Leader
Dr. William S. Farrell, Jr., Test Team Leader
Dr. John Caylor, Test Team Leader
Dr. Louis A. Armijo, Field Coordinator
Ms. Marie R. Peirano, Field Coordinator
Ms. Jeanette D. Wheeler, Test Editor and Production Coordinator

Mr. Steven Zwillinger, Department of Education Project Officer, provided support to staff throughout the project.

INTRODUCTION

Introduction

This module, Module 20, is the last in a series of four designed to help vocational educators develop and use vocational competency measures. Module 17 provided an overview of using competency measures in vocational education programs. Module 18 discussed how to determine requirements for vocational competency measures. And Module 19 presented a step-by-step approach to developing the competency tests.

The purpose of this module is to help you validate a competency test after it has been developed; to determine how the test results will be reported; and to consider ways of setting standards for passing or failing the test. The discussion presented here is based on the experiences of the American Institutes for Research in conducting the Vocational Competency Measures (VCM) project for the U.S. Department of Education as well as on prior test development experience of project staff.

Overview

The first section of the module discusses possible approaches for determining the content validity of a competency test. As an illustration, it describes the approach AIR used in the Vocational Competency Measures project. It also comments on predictive validity--the ability of the test to predict job success--and maintaining test validity over time once a test has been developed and validated.

The next section of the module considers possible uses for vocational competency tests and how test results can be reported in accordance with the intended use. It discusses four possible ways of reporting results and presents other important aspects of reporting test scores: whether scores should be reported on a group or an individual basis, and to whom they should be reported.

The module concludes with the controversial question of how to set standards for passing or failing a test. It presents several approaches to setting standards for vocational competency tests and highlights important points to remember: standards must be determined on a reasonable basis, defensible from both a technical (psychometric) and a legal standpoint, equitably applied across all examinees, and acceptable by the users of the results of the test.

Instructions to the Learner

The Self-Check items and possible responses to them are found in the Appendices. These questions have two purposes. First, before you begin work on the module, you may use them to check quickly whether you have already learned the information in previous classes or readings. In some instances, with the consent of your instructor, you might decide to skip a whole module or parts of one. The second purpose of the Self-Check is to help you review the content of modules you have studied in order to assess whether you have achieved the module's goals and objectives.

You can also use the list of goals and objectives that follows to determine whether the module content is new to you and requires in-depth study, or whether the module can serve as a brief review before you continue to the next module.

Goals and Objectives

Goal 1: Explain the procedure for determining the content validity of a vocational competency test.

Objective 1.1 State the purpose of determining the content validity of a test.

Objective 1.2 Compare content validity and predictive validity of a test.

Objective 1.3 Describe the process of maintaining test validity.

Goal 2: Summarize the possible uses for vocational competency tests and ways of reporting test results.

Objective 2.1 List three possible uses for vocational competency tests.

Objective 2.2 List three possible ways of reporting test results.

Objective 2.3 State the basis for determining whether test scores should be reported on a group or an individual basis.

Objective 2.4 State the basis for determining who should receive test scores.

Goal 3: Explain how to set standards for passing or failing a vocational competency test.

Objective 3.1 Describe one approach to setting standards for vocational competency tests.

Objective 3.2 List the most important considerations in setting test standards.

Resources

In order to complete the learning activities in this module, you will need information contained in the following publication:

Erickson, R. C., & Wentling, T. L. Measuring student growth: Techniques and procedures for occupational education. Urbana, Ill.: Griffon Press, 1976.

GOAL 1: Explain the procedure for determining the content validity of a vocational competency test.

Validating the Test

After a test has been developed, regardless of the purpose for which it is intended, an effort must be undertaken to determine its validity. In simplest terms, the validity of a test provides an indication of whether and to what extent the test will be useful for its intended purpose. Does the test, in fact, provide the sort of information it is supposed to provide?

For tests that are intended primarily to predict future behavior or performance, the preferred way to determine a test's validity is to correlate test scores with measures of actual performance at a later date. This is termed predictive validity. By its nature this procedure requires a fairly long period of time. First, the test is given and then a period of time must elapse that is long enough to enable the individuals involved to have an opportunity to demonstrate how well they can do in all aspects of their chosen career in an actual work situation. Of course, it would be possible to locate individuals already established in their careers, obtain ratings of their on-the-job performance, administer the test to them, and determine the relationship between the two measures--this would be termed concurrent validity. A third procedure that is particularly relevant for competency tests requires the use of experts to assess how representative the coverage in the test is of the area in question--this is termed content validity.

Content Validity

The determination of content validity is based on expert judgments rather than statistical procedures. Thus a test's level of content validity is expressed in qualitative rather than quantitative terms. (A test might be spoken of as having a high, medium, or low level of content validity.)

To determine the content validity of a test, a group of experts in the content area covered by the test is asked to review the actual content of the test and to provide an indication of the degree to which the test content covers the material that should be covered given the purpose of the test.

For example, assume that a test has been developed to evaluate the proficiency of individuals completing a week-long course on developing dental x-rays. The test covers both knowledge and performance aspects of the course. A group of content experts, probably consisting of persons who teach such courses and of persons who actually develop dental x-rays, would be asked to review the content of the new test and, based on their expert knowledge, provide an evaluation of how well the questions and problems presented in the test cover the skills and abilities that should be possessed by individuals completing the course.

The actual evaluation of the content of the test could be carried out in any of several ways. Among these are the following:

- Provide the content experts with copies of the test and ask each of them to make an independent, global evaluation.
- Provide the content experts with copies of the test and ask them to arrive at a consensual, global evaluation.
- Provide the content experts with copies of the test and ask them to provide independent evaluations on the relevance of each item in the test, and whether any major areas of significance were omitted.
- Provide the content experts with copies of the test and ask them to arrive at a consensual evaluation for each item, and whether any major areas of significance were omitted.
- Follow any of the four strategies described above, but provide the content experts with copies of the test outline rather than the actual test.
- Ask the content experts to first develop, either individually or collectively, an outline of what should be covered in a test for a given purpose and then ask them to evaluate the new test with regard to that outline, either individually or collectively.
- Provide the content experts with an exhaustive test outline, developed either by the test developer or some third party, of all the topics that could be covered in the test and then ask them to evaluate the new test with regard to that outline, either individually or collectively.

Obviously there isn't any set procedure for determining the content validity of a test. The possible approaches vary greatly with respect to the amount of time required of the content experts, the types of tasks to be performed by the content experts, and the nature of the results produced. The exact procedure to be followed in any particular case would, of course, depend on the time and funds available, and the wishes of the test developers.

While many other formats are possible, Figure 1 on the next pages illustrates the way in which information on the content validity of a test was collected in the AIR Vocational Competency Measures project. In this instance, the content experts were provided with outlines of the content covered in the paper-and-pencil sections and a listing of the individual performance tasks, and were asked to rate each element of the outline on a four-point scale of importance. It should be noted that the outlines reflect the comprehensive requirements for each area as determined by the procedures discussed in Module 18 of this series.

Clearly, if the content experts rated all the elements of the outline as very important, or even a mix of very important and fairly important, the test developers could be confident that they had included relevant material in the test. However, this approach does not indicate whether the content experts feel that all of the most important topics have been included but only that the topics that have been included in the test are important. For this reason, each rater should also be asked to list any areas that were omitted from the test that are of major importance.

If appropriate procedures were used in the initial determination of the topics to be included in the test, a step that also required the assistance of content experts, then more complex approaches to content validation are not required. The post-test development, content validation step in effect becomes a verification of the results of the step in which the test content was initially determined.

The items or performance measures included in a test usually do not cover the complete domain of the content area of the test. The primary reason for this lack of total coverage is the fact that tests must be limited in length and in the time they require for administration. As a consequence, test content coverage is typically limited to those topics considered to be most important.

COMPETENCY TEST CHECKLIST FOR
ELECTRONICS TECHNICIAN

Certain knowledges and skills are expected when hiring someone who has just completed a training program. The American Institutes for Research has been developing a test for Electronics Technician to measure these skills and knowledges. We are interested in reactions of employers and supervisors to the proposed content areas. We would like to know how important you think it is for an electronics technician, who has recently completed a training program, to know and be able to perform certain tasks which are measured by the test. The test is in two sections--a performance, hands-on section and a job knowledge, paper-and-pencil section.

Please indicate how important you think each test content area is, by circling one number in each row to the left of each test content area statement on this and the next page.

Of No Importance	Of Minor Importance	Fairly Important	Very Important	OUTLINE OF TEST CONTENT AREAS		X
				PERFORMANCE, HANDS-ON SECTION		
1	2	3	4	a. Replace components on PC board	a.	____
1	2	3	4	b. Measure voltages and optimize bias of two-supply CE amplifier	b.	____
1	2	3	4	c. Assemble CE amplifier and analyze distortion	c.	____
1	2	3	4	d. Identify circuit components from schematic and from equipment	d.	____
1	2	3	4	e. Measure voltages and calculate gain in common base amplifier	e.	____
1	2	3	4	f. Assemble and test bridge-type power supply	f.	____
1	2	3	4	g. Establish feedback and determine gain of Op-Amp (analog)	g.	____
1	2	3	4	h. Assemble and test monostable multivibrator	h.	____
1	2	3	4	i. Assemble and test IC digital clock pulse circuit	i.	____
1	2	3	4	j. Determine frequency response of Op-Amp	j.	____
1	2	3	4	k. Analyze operation of simplified differential amplifier	k.	____
1	2	3	4	l. Assemble and test Mod 5 shift counter	l.	____

FIGURE 1. Example of approach used in VCM project for determining the content validity of a test.

Of No Importance	Of Minor Importance	Fairly Important	Very Important	OUTLINE OF TEST CONTENT AREAS
				JOB KNOWLEDGE, PAPER-AND-PENCIL SECTION
1	2	3	4	A. Using General Purpose Test Equipment (oscilloscopes volt-ohmmeters, function generators, frequency counters, power supplies, etc.)
				B. Using Hand Tools
1	2	3	4	1. Basic hand tools
1	2	3	4	2. Soldering tools
1	2	3	4	3. Alignment tools
				C. Troubleshooting
1	2	3	4	1. Isolating and identifying faults
1	2	3	4	2. Analyzing circuit measurements (analog or digital)
1	2	3	4	3. Analyzing circuit functions (analog or digital)
1	2	3	4	D. Selecting and Replacing Components
				E. Fabricating Electronic Equipment
1	2	3	4	1. Identifying and select components
1	2	3	4	2. Using appropriate types of solder
1	2	3	4	3. Applying basic electronic construction techniques
1	2	3	4	F. Calibrating Electronic Equipment
				G. General Knowledge and Procedures
1	2	3	4	1. Performing math and electronics calculations
1	2	3	4	2. Reading schematics
1	2	3	4	3. Using specified test procedures
1	2	3	4	4. Interpreting test results

FIGURE 1 (continued)

Predictive Validity

As was pointed out earlier, some sorts of tests are developed primarily with the hope of predicting some future performance or behavior. For example, an end-of-course test might be intended to predict success on the job during the first six months after leaving the course, or a test might be developed to be given to applicants for a particular course with the hope of predicting performance in the course. In such situations, it is appropriate to determine the statistical or criterion-related validity of the test as well as the content validity. The criterion-related validity of a test is usually expressed in terms of the correlation coefficient between scores on the test and some performance measure. Among the possible criterion measures for a test to predict job success at the end of six months are:

- Promotion or not
- Productivity measure
- Error rate
- Supervisor's rating

It should be noted that there is no single, universal criterion measure against which a test can be statistically validated. The criterion to be used in any particular instance will depend on the purpose of the test and the nature of the available criteria. It is also important to note that all of these criteria, as well as almost all others that might be thought of, require the collection of both test scores and performance measures on an adequate number of examinees. Such an effort can be very time-consuming and expensive, and may well require the cooperation of many employers.

A statistical validation study should not be undertaken at all lightly, and it is recommended that such a study not be undertaken unless an individual with a background in psychometrics is available to guide the effort. Test developers interested in conducting a statistical validation are also urged to consult one of the many books on selection and classification or employee testing. (See the Recommended References for a partial listing of some of the important ones.) Of particular value, though it is somewhat technical, is a book called Personnel Selection by Robert L. Thorndike (1949).

Maintaining Test Validity

At first glance, it may appear unnecessary to worry about maintaining the validity of a test that has just been developed and validated. However, time does pass and things do change with time. For example, it wasn't very long ago that auto mechanics did not have to know about electronic ignitions and hospital x-ray technicians did not have to know about CAT scanners. Because of changes like these, vocational programs must also change to some degree every year and during some years there are large changes. Thus, every few years it is necessary to review the content of competency tests to make certain that they don't include topics that are no longer important and that they have not missed topics that have recently become important.

The best way to carry out a validity check is to rely on your employer advisory committee and, on a regular basis, ask them to review the test content, looking for topics that should be deleted or added, or whose emphasis should be changed. If there are required changes, then either the changes should be made or the test should be removed from use.

On a periodic basis, depending on the changes in job content that are taking place, you should plan a more comprehensive survey of job requirements as was done for developing the original set of test requirements (see Module 18). Only if such continuing review is carried out can you be sure that the test is still serving the purpose for which it was intended.

Individual Study Activities

1. The determination of the content validity of a test can be carried out in any of several ways. This module describes the approach AIR used in the Vocational Competency Measures project. Obtain a competency test that you have used or with which you are familiar. Contact the developer or publisher of the test and conduct an interview to determine the process that was used in validating that test. Write a report of your findings and share them with the class.

2. Select a reference of your own choosing on competency test development. Read a chapter or section pertaining to content validation approaches. Select an approach that you would find useful in your particular setting and briefly describe it in a short paper. Provide reasons for your selection of that approach to content validation.

Discussion Questions

1. "There are very few standardized instruments that can be considered perfectly valid. Few, if any, provide complete measurements of that which they were designed to measure--nothing more or nothing less" (Erickson & Wentling, 1976, p. 309). How much validity is enough? How much validity should a standardized instrument have? Discuss these questions in class and see if you can arrive at some general guidelines regarding validity when selecting standardized instruments for occupational programs.

2. When selecting a standardized instrument for occupational programs, why is it important to consider the predictive validity of the test? What information will predictive validity provide that content validity does not? What is the basic difference in the processes for determining predictive validity and content validity?

Group Activity

1. Divide the class into small groups (4-5 people). Each group will meet separately to develop a plan for maintaining validity of a locally-developed vocational competency test. When the groups reconvene, each should present its plan to the class. When all the plans have been presented, note the similarities and differences of the plans. See if the class can come up with one plan that incorporates the best features of all the small group plans.

GOAL 2: Summarize the possible uses for vocational competency tests and ways of reporting test results.

Using Test Results

Once a test has been developed, what is to be done with it? How are the test results to be used? These are really questions that must be addressed early in the test development process, as noted in the first module of this series. However, since test usage has a great deal to do with how test results will be reported, this topic will be discussed briefly here. Several possible uses for vocational competency tests are obvious:

- As a course final exam
- As a vehicle for professional certification
- For comparing the collective performance of individuals from different schools
- A diagnostic test to be used to determine areas in which students need more work
- For job selection
- For job assignment or classification
- For determining areas in which different schools are weak and thus need to change their programs

The intended use of the test will play a major role in determining how the test results will be reported.

Reporting Test Scores

Based on the possible uses to which the test results might be put, as listed in the section above, at least four different ways of reporting results are possible. These four ways, listed in order by the amount of information they provide, are:

- a single pass-fail mark for the entire test
- a single numeric score for the entire test

- a pass-fail mark for each independent part, or content area, of the test
- a numeric score for each independent part, or content area, of the test

Norm-referenced vs. criterion-referenced tests. Although the distinction is by no means complete, it is often useful to think of tests, and thus test score reporting approaches, as falling into two classes: norm-referenced tests and criterion-referenced tests.

In simple terms, a norm-referenced test may be thought of as a test on which examinees are compared with each other and their results are reported in terms of their standing with regard to some standard reference group, the norm group. A criterion-referenced test may be thought of as a test on which examinees are compared with some preset, external, hopefully objective standard, the criterion.

Results on norm-referenced tests can usually take a large range of possible values. On the other hand, results on criterion-referenced tests often, but certainly not always, take only one of two possible values: pass when the examinee meets or exceeds the preset standard, and fail when the examinee fails to meet the preset standard. There are also instances when results on criterion-referenced tests, especially for criterion-referenced tests that attempt to measure more than one criterion or objective, are reported in terms of the number or percent of items answered correctly.

Choosing the numeric scale. If scores are to be reported on a numeric scale they should, of course, be related in some way to the number of questions that the examinee answered correctly, or the number of elements of the performance problem or problems that were carried out correctly. These numeric scores could be reported in at least three different ways: (1) the raw score (number of questions correct) on the test, or test part; (2) the percent of questions correct; or (3) some form of standardized score (for example, in terms of a distribution of scores with a mean score of 50 and a standard deviation of 10).

The selection of a scale on which to report test scores should depend, at least in part, on the scores that are to be reported. For example, if only a single test score is to be reported for the entire test, then it makes very little difference which of the three types of scales is used, since the results will be comparable in each case. However, if a numeric score is to be reported for each of several parts or sections

of the test, then it makes sense to use a score reporting scale in which the same score means the same thing regardless of the part of the test to which it applies. A raw score of seven means one thing if there are seven questions on the test part, but it means something very different if there are 15 questions on one part and 30 on another. On the other hand, a score of 69% correct, or a standard score of 55 (on a scale with a mean of 50 and a standard deviation of 10) are more consistent in meaning regardless of the number of questions included in the test part. In general, the test score reporting format selected should be one that reports useful information in an easily understood way.

There are no rules that govern whether converted scores should be reported in terms of percentiles or standard scores, since each has different strengths and weaknesses. To help you decide which score reporting system will be most appropriate for your purposes, a summary of the advantages and disadvantages of each approach is presented below (adapted from Cronbach, 1960, p. 86-87).

Percentile scores--

Advantages:

- Easily understood by persons without statistical training
- Easily computed
- May be interpreted exactly even when the distribution of test scores isn't normal

Disadvantages:

- Magnify small differences in score near the mean and minimize large differences in score near the end of the distribution
- May not be used in many statistical calculations

Standard scores--

Advantages:

- Differences in standard scores are proportional to differences in raw scores
- Appropriate to use in statistical calculations

Disadvantages:

- Cannot be interpreted readily when distributions are skewed
- Often difficult for untrained persons to understand

Generally, statisticians prefer standard scores and laypersons prefer percentiles.

Two other important aspects of reporting test scores are: (1) whether scores should be reported on a group or an individual basis, and (2) to whom they should be reported. As with so many other questions related to test development, these questions should be answered on the basis of the intended use of the test.

Individual versus group reporting. If the test is intended primarily for the use of schools as a course improvement aid, then combining the results for all the individuals in a given school, course, or class is the logical approach, since this will provide information in an immediately useful form. However, if the purpose of the test is to provide information on the performance of individual examinees, then scores must be reported for individuals. In addition, it would probably be useful to report school, course, or class averages since most instructors will want this information.

Even in cases where the purpose of a test is to provide group as opposed to individual data, it is still a good idea to also provide the individuals with copies of their own results. Examinees generally want to know how they do on a test and knowing that they will get their own results may help to motivate the examinees to try to do well. Clearly, the test results should not be reported to outside individuals (for example, potential employers) without the permission of the examinee.

Another point to remember is that when tests are administered for the purpose of evaluating a program rather than individuals, it may not be necessary to administer every test, or test part, to every individual. This is especially true when a fairly large number of individuals will be tested. For example, consider a program with a total of 100 students, which is to be evaluated with four performance measures. The students could be randomly divided into four groups with each group taking one performance measure, or divided into two groups with each group taking two performance measures. Such an approach can greatly reduce the testing time while still providing vital information. In general, if a group is going to be divided so different subgroups of individuals take different performance measures, each subgroup should consist of at least 15 to 20 individuals.

Who should receive test scores. Are the test scores to be reported to the individual examinees, or only to the instructors? In general, if scores are to be reported back to

the individual examinees, and we strongly recommend that in most cases the examinees should be provided with feedback on their performance, it is best to prepare individual score reporting forms that give a detailed explanation of what the scores mean and do not mean. Such detailed score reporting forms are often not needed if scores are only to be used by instructors, but here care must be taken to see that the instructors do in fact understand what the scores mean.

The question of who should receive test scores can also be an issue when average scores for a school, course, or class are to be calculated. For example, should the scores for all the schools, courses, or classes be reported to everyone; or should an individual school, course, or class receive only its own scores along with the mean and standard deviation of all the schools, courses, or classes combined? In most instances, it is best not to report all scores to all parties since this often can lead to "I'm better than you are" situations.

Remember--tests are not perfect. A final consideration in reporting test scores is the fact that while a test score can be considered an estimate of where the examinee stands on the dimension underlying the test, a single test score should not be considered a definitive measure of the examinee's true standing. This is the case since all test scores contain an error component, and for any given individual the size of this error component is unknown. However, based on a group of examinees, it is possible to develop an estimate of the accuracy of test scores known as the standard error of measurement.

While a full discussion of the standard error of measurement is beyond the scope of this module, it is an important statistic and deserves at least a brief mention. An individual may be thought of as having a certain amount of the ability or characteristic which underlies a test. Because all actual test scores are subject to error, the score obtained by an individual on a test may or may not accurately reflect that individual's true ability. If an individual were to take a whole series of equivalent tests, we would expect the distribution of obtained test scores to cluster around a score representing the true underlying ability. This distribution of test scores for an individual could be expected to take the form of a normal distribution. The standard error of measurement for this distrubution is a measure of the degree of variability of the test scores. Since it is not practical to administer many test forms to a single individual, the standard error of measurement is in fact calculated using the reliability of the test and the standard deviation of test scores for the group on which the reliability is calculated. Once the standard error of measurement has been calculated,

given an obtained test score for an examinee, we can be approximately 68% sure that the examinee's true score (a measure of true ability on the underlying characteristic) lies within plus or minus one standard error of measurement, and approximately 95% sure that the examinee's true score lies within plus or minus two standard errors of measurement. For example, if a test had a standard error of measurement of 3 and an examinee received a score of 25, we could be 68% sure that the individual's true score is between 22 and 28 (inclusive) and 95% sure it is between 19 and 31 (inclusive). The most important point for you, as a test developer or user, to keep in mind is that test scores are subject to error, sometimes large errors, and do not necessarily represent the examinee's true ability.

Individual Study Activities

1. Write your definitions of the following terms:

   (a) norm-referenced test
   (b) criterion-referenced test
   (c) raw score
   (d) mean score
   (e) standard deviation

   You may want to use a standard text on testing or the Recommended References listed for this module to help you arrive at your definitions.

2. Select a vocational course or program with which you are familiar that uses vocational competency tests. Identify the ways in which these tests are used. Then for each use, identify the way in which the test results are reported. Present your findings to the class.

Discussion Questions

1. "The most important aspect of scoring students' performance on measurement instruments is accuracy" (Erickson & Wentling, 1976, p. 351). Provide reasons to support this statement. How does the "standard error of measurement" contribute to ensuring accuracy of test scores?

2. "Just as the philosophical bases for criterion-referenced measurement differ from norm-referenced measurement, so do some of the methods of scoring, reporting, and interpreting the measurement results obtained with these two approaches to measurement". (Erickson & Wentling, 1976, p. 399). What are some of the ways in which these methods differ? List them on the chalkboard.

Group Activity

1. Have the class break into three groups. Each group will interview vocational education officials in a school district within the state--one group will select a small district, another group a medium-size district, and the third group a large school district. Each group will conduct interviews by phone or on-site to determine district policy on vocational competency testing. How are competency tests used in these districts and how are test results reported? Each group should make a report on its findings at the next class session. Compare the findings among the three sizes of school districts.

GOAL 3: Explain how to set standards for passing or failing a vocational competency test.

Setting Test Standards

Few issues in testing have generated so much smoke and so little fire as the question of how to set standards for passing or failing a test. There is at present no absolute rule that can be used to set standards. As a result, arbitrary rules have often been used. A typical arbitrary rule used by many classroom teachers goes something like this: "I teach the course and I know how to write a test. Any student who gets from 91% to 100% of the questions right gets an A; any student who gets from 81% to 90% right gets a B; etc." While it is an easy rule to state, it does require some significant assumptions about the teacher's real knowledge of how an examinee's performance reflects his or her knowledge or ability, and about the teacher's ability to construct tests with known characteristics.

Other arbitrary rules assume that a certain proportion of a class will earn an A; a certain proportion, a B; etc. This is, of course, known as grading on the curve; and here, too, an important assumption is made about the distribution of ability on the dimension underlying the test. Most commonly, this assumption is that ability is normally distributed and thus there should be more Cs than Bs, more Bs than As, as many Ds as Bs, and as many Fs as As. While such an assumption may be justified when dealing with a randomly selected group in which none of the individuals has had any experience or training relative to the test content, it is probably not a good assumption when prior training and/or screening have taken place.

Another fairly commonly used rule for criterion-referenced tests, based on objectives and where scores are reported for classes or schools, is that if 80% of the students get at least 80% of the questions correct, then the class or school will be considered to have mastered the objective. Like the other rules just discussed, this too is entirely arbitrary and there is nothing magic about it. It only seems to be concrete because it is stated in quantitative terms.

Approaches to Setting Standards for Vocational Competency Tests

If the test you are developing is one on which individuals will receive a passing or failing grade, you should plan to use a logical, although it will probably still be arbitrary, approach to setting standards. In order to do this, you are strongly urged to work with the advisory committee you set up early in the test development process. Seek their advice, help, and cooperation. For example, suppose you are developing a test for auto mechanics. Several members of your test advisory committee probably employ auto mechanics. Ask these individuals to identify, and then let you administer the test to, their recently hired employees whom they consider minimally competent as auto mechanics. If a satisfactory number of such minimally competent individuals can be found and tested, then their test results can be used to set the standard that defines a minimally competent person. Note that special attention should be given to making sure that the minimally competent, rather than the best individuals, are tested, since to test only the best individuals would result in setting the standards too high. Even if the members of your test advisory committee do not have enough minimally competent employees, they may be able to provide you with the names of, and an introduction to, other employers whose employees could be included.

Another possible approach to setting standards for vocational competency tests is based on the fact that many instructors in vocational programs have worked or still do work in the field in which they are instructing. As a result, these individuals probably have a good idea of what it will take to do a good job once a student finishes the program. Ask these instructors to select individuals completing the program whom they consider to be minimally competent for an individual just entering the field. Again, be sure they do not nominate the best students in the class. Use the results for this minimally competent group of students to set the minimum passing standard.

A third approach, though it is not as satisfactory as either of the preceding two, is to have the members of the test advisory committee meet to go over the final content of the test and, based on this content review, to set the standards. While this approach is arbitrary, it does have the advantage that it is based on judgments of several persons who are knowledgeable about the field; and it is likely to be accepted by the user community since the test advisory committee members are from the field.

Impact of Errors in Test Standards

Aside from the technical/mathematical considerations that should be taken into account when setting test standards, an important human (and legal) concern must be the consequences of errors in cut-off scores. Clearly, we would like to set cut-off scores so all qualified individuals pass, while all nonqualified individuals do not pass. However, such perfect cutting points do not exist. So, is it best to set the cut-off scores high so that not only do virtually all nonqualified individuals fall below it but also a fair number of qualified individuals as well? Or is it best to set the cut-off score low so virtually all qualified and more than a few nonqualified individuals fall above it? Which of these approaches (or some intermediate approach) is taken should depend on the consequences of a classification error. To incorrectly classify a would-be physician as qualified to practice is far more serious than to incorrectly classify a first-year student as qualified to take a second-year course. We should aim to set cut-off scores so as to minimize the total harm (to individuals and society) that will result from classification errors.

Keeping Test Standards Up-to-Date

Once test score standards, or cut-off scores, have been set, they should not be considered as fixed and invariant for the life of the test. Instead, such test standards should be reviewed on an ongoing basis and revised or adjusted whenever necessary. Such revision may be necessary because of such factors as: changes in job content, changes in course content, changes in employer expectations as to what constitutes minimum competence, or even the discovery that the initial standards were set incorrectly.

Legal Considerations in Setting Performance Standards

Technical issues are not the only problems that must be faced in setting performance standards. Schools are more and more facing the threat of legal action related to these standards. Tractenberg (no date), in his overview of the legal implications of performance testing in vocational education, stresses that these legal concerns "should play a significant role in the development of performance testing" (p. 96). Concerns that relate explicitly to setting test-score standards include:

- the number of proficiency standards that will be set,
- the level(s) at which these standards will be set,
- whether the standards will be for school programs or students, and
- the consequences for failing to achieve the standards.

Tractenberg recommends two courses of action. In terms of level at which proficiency standards are set, he suggests "as a practical matter, unless a particular program is specifically designed to equip its students for journeyman positions, the standards should be geared to entry-level positions. The more important issue is likely to be whether the standards actually relate to the marketplace" (p. 101). In his discussions of the consequences of failing to achieve the test standards, Tractenberg recommends that:

> The preferable, and in some cases the required, response to evidence that particular students had failed to meet proficiency is to direct appropriate educational assistance to them. This may take the form of remediation for the individual students; it may involve broader programmatic or personnel responses. Surely, if a substantial percentage of the school's or program's students is failing to meet statewide or local standards, the overall educational program, including the quality of instructional staff, should be evaluated and perhaps upgraded (p. 102).

Pullin (no date, p. 118) raises yet another legal issue relative to test scores. In her consideration of privacy and confidentiality in performance testing, she recommends that:

- Test scores should not be disclosed to persons outside the school or to those not directly involved with the student's training without consent.
- Test scores should not be divulged to potential employers without the written consent of the parent, or if the student is over 18, the student.
- Interpretation of test results should be made available to students' parents.
- Tests should not include questions that unnecessarily infringe on students' privacy.

Tractenberg (no date, p. 103), looking toward future developments in the area of legal issues surrounding performance testing in vocational education, makes the following important recommendation:

> Vocational educators should not simply sit back and wait to be sued. They should deal in some preventive maintenance--they should attempt to head off legal challenges by fashioning and implementing performance testing programs in the most careful manner possible. If they do so, the law and the courts will have been an important partner in educational and professional reform.

Points to Remember About Test Standards

The most important points to remember about setting standards are that they must be determined on a reasonable basis, defensible from both a technical (psychometric) and a legal standpoint, equitably applied across all examinees, and acceptable by the users of the results of the test.

Individual Study Activities

1. Using the Recommended References listed in this module or a resource of your own choosing, make a list of techniques for setting test standards. Which of these techniques apply to norm-referenced tests and which apply to criterion-referenced tests? Note any similarities and differences.

2. Few issues in testing have generated so much smoke and so little fire as the question of how to set standards for passing or failing a test. Conduct a literature review of recent journal articles that discuss this issue. Summarize the points of view presented in these articles. Then summarize your point of view on setting test standards upon completion of your readings.

Discussion Questions

1. Grading on the normal curve assumes that student achievement is a normally distributed trait among the students in the classes in which the system is used. This assumption is not always valid, particularly when a class is composed of gifted, handicapped, or disadvantaged students. What are ways of setting test standards that accommodate individuals with special needs?

2. Criterion-referenced scores are often of the pass-fail type. The criterion or minimum level for a passing performance on an achievement test is established prior to the administration of the test, and generally prior to the instruction that is covered by the test. However, once the pass-fail grade is recorded, much valuable information is lost. For example, students who can accurately type at 58 words per minute and at 30 words per minute may each receive the same failing grade (adapted from Erickson & Wentling, 1976, p. 401). What are some examples where such a loss of precision is tolerable? What are some examples where such a loss of precision would <u>not</u> be tolerable?

Group Activity

1. Break the class into small groups (3-4 people). Each group should meet and select an approach to setting standards for vocational competency tests. When the class reconvenes, each group should present its approach and support it by describing its advantages.

SUMMARY

Summary

Any test, once it has been developed and regardless of its purpose, should be validated to ensure that it will be useful for its intended purpose. This is true for vocational competency tests, and the AIR approach in the Vocational Competency Measures project provides a useful model. Continuing review is necessary to be sure that the test is still serving the purpose for which it was intended.

The intended use of the test, which is determined early in the test development process, plays a major role in determining how the test results will be reported. A variety of ways is possible. Whether scores should be reported on a group or on an individual basis, and to whom they should be reported are also questions that need to be answered on the basis of the intended use of the test.

Another critical issue in testing is how to set standards for passing or failing a test. There is at present no rule that can be used universally to set standards, and all existing approaches are to some degree arbitrary. However, it is important to be familiar with these approaches and to recognize their strengths and weaknesses. The most important points to remember about setting standards are that they must be determined on a reasonable basis, defensible from both a technical (psychometric) and a legal standpoint, equitably applied across all examinees, and acceptable by the users of the results of the test.

APPENDICES

Self-Check

GOAL 1

1. What is the purpose of determining the content validity of a test?

2. What is the purpose of determining the predictive validity of a test?

3. Briefly describe the process of maintaining test validity.

GOAL 2

1. What are three possible uses for vocational competency tests?

2. What are three possible ways of reporting test results?

3. What is the basis for determining whether test scores should be reported on a group or an individual basis?

4. What is the basis for determining who should receive test scores?

GOAL 3

1. Describe one approach to setting standards for vocational competency tests.

2. What are the most important considerations in setting test standards?

Self-Check Responses

GOAL 1

1. The purpose of determining the content validity or relevance of a test is to determine whether and to what extent the test content covers the material that should be covered.

2. The purpose of determining the predictive validity of a test is to determine the ability of the test to predict some future performance or behavior of an individual.

3. Every few years it is necessary to review the content of tests to make certain that they don't include topics that are no longer important and that topics that have recently become important are included. The best way to carry out this process is to rely on the employer advisory committee for the test development process and, on a regular basis, ask them to review the test content, looking for topics that should be deleted or added, or whose emphasis should be changed. If there are required changes, then either the changes should be made or the test should be removed from use.

GOAL 2

1. A vocational competency test may be used as a course final exam; for certification purposes; for comparing the collective performance of individuals from different schools; as a diagnostic test to determine areas in which students need more work; for job selection; for job assignment or classification; as a diagnostic test to be used to determine areas in which different schools are weak and thus need to change their programs.

2. Test results may be reported as a single pass-fail mark for an entire test; a single numeric score for an entire test; a pass-fail mark for each independent part, or content area, of the test; a numeric score for each independent part, or content area, of the test.

3. If the test is intended primarily for the use of schools as a course improvement aid, then the logical approach is to report test scores on a group basis. If the purpose of the test is to provide information on the performance of individual examinees, then scores must be reported for

individuals. The intended purpose of the test is the basis for determining how to report test scores.

4. Again, the intended use of the test is the basis for determining who should receive test scores. In most if not all cases, examinees should be provided feedback on their test performance. Special care should be taken to ensure that confidentiality and examinee privacy are protected.

GOAL 3

1. Approaches to setting standards for vocational competency tests include: having recently hired employees whom an advisory committee considers minimally competent take the test and use their performance to set the standard that defines a minimally competent person in a specific occupation; having instructors select students completing a program whom they consider minimally competent for entering the field and use their test results to set the standard; having the test advisory committee review the final content of the test and use their judgment to set standards.

2. The most important considerations in setting test standards are that they are determined on a reasonable basis, defensible from both a technical (psychometric) and a legal standpoint, equitably applied across all examinees, and acceptable by the users of the results of the test.

Recommended References

Cronbach, L. J. Essentials of psychological testing (2nd ed.). New York: Harper & Row, 1960.

Dunnette, M. D. (Ed.). Handbook of industrial psychology. Chicago: Rand McNally College Publishing Co., 1976.

Dunnette, M. D. Personnel selection and placement. Chicago: Brooks-Cole, 1966.

Erickson, R. C., & Wentling, T. L. Measuring student growth: Techniques and procedures for occupational education. Urbana, Ill.: Griffon Press, 1976.

Guion, R. M. Personnel testing. New York: McGraw Hill, 1965.

McCormick, E. J., & Tiftis, J. Industrial psychology (6th ed.). Englewood Cliffs, NJ: Prentice-Hall, 1974.

Pullin, D. Performance testing in vocational education--Lessons to be learned from the minimum competency testing movement. In J. E. Spirer (Ed.), Performance testing: Issues facing vocational education. Columbus, Ohio: Ohio State University, National Center for Research in Vocational Education, n.d.

Thorndike, R. L. (Ed.). Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Thorndike, R. L. Personnel selection. New York: John Wiley & Sons, Inc., 1949.

Tractenberg, P. L. Legal implications of performance testing in vocational education: An overview. In J. E. Spirer (Ed.), Performance testing: Issues facing vocational education. Columbus, Ohio: Ohio State University, National Center for Research in Vocational Education, n.d.

Wentling, T. L. Evaluating occupational education and training programs (2nd ed). Boston: Allyn and Bacon, 1980.

Wood, D. A. Test construction: Development and interpretation of achievement tests. Columbus, Ohio: Charles E. Merrill, 1961.