DOCUMENT RESUME

ED 052 259 TM 000 769

AUTHOR	Boyd, Joseph L., Jr.; Shimberg, Benjamin
TITLE	Developing Performance Tests for Classroom Evaluation.
INSTITUTION	ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, N.J.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	TMR-4
PUB DATE	Jun 71
NOTE	17p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Aptitude Tests, Behavioral Objectives, Evaluation Criteria, *Evaluation Techniques, Measurement Techniques, *Performance Criteria, *Performance Tests, Rating Scales, Scoring, Task Performance, *Test Construction, Test Validity, Vocational Aptitude

ABSTRACT

Performance tests are considered significant instruments in facilitating the accurate assessment of an individual's overall competency. The importance of measuring performance and the advantages of using performance tests are discussed. Developing the test involves identifying the objectives to be measured, the task to be performed, and the time required for its performance. Once the procedures, directions, equipment, and scoring methods have been developed, they should be formalized through the preparation of three documents: Instructions to the Administrator, Instructions to the Examinees, and Rating and Scoring Form. Finally, the factors involved in grading the performance test are detailed, including a discussion of the relative importance of product and process evaluations. (PR)

TM REPORTS

NUMBER 4

Developing Performance Tests for Classroom Evaluation

ERIC CLEARINGHOUSE ON TESTS, MEASUREMENT, & EVALUATION ■ EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY 08540
Conducted by Educational Testing Service in Association with Rutgers University Graduate School of Education

The Clearinghouse operates under contract with the U. S. Department of Health, Education and Welfare, Office of Education. Contractors are encouraged to express freely their judgment in professional and technical matters. Points of view expressed within do not necessarily, therefore, represent the opinions or policy of any agency of the United States Government.

June 1971

DEVELOPING PERFORMANCE TESTS

FOR

CLASSROOM EVALUATION

Joseph L. Boyd, Jr.

&

Benjamin Shimberg

ERIC Clearinghouse on Tests, Measurement, and Evaluation

INTRODUCTION

Since World War I, testing specialists have made great advances in measuring both aptitude and achievement by means of paper-and-pencil tests. A multimillion dollar industry has been built on the ability of such tests to predict with reasonable accuracy the future success of students in elementary and secondary schools, in college, and in professional schools. Written tests have also been used to screen applicants for training programs and to identify workers who are most likely to be successful on the job.

By contrast, except for the business field, where typing and shorthand tests are widely used, performance testing has been largely neglected. The efficiency and economy of paper-and-pencil testing made it attractive to educators and to employers--written tests could be given to large groups at a single sitting, and the answer sheets could be scored quickly, either by hand or electronically. Performance tests, on the other hand, almost always have to be given on an individual basis. Such tests take longer to administer, and the scoring cannot be done as quickly or as economically.

The range of applications for performance tests is very large. Typically, shop and home economics teachers have used performance measures to evaluate the skill development of students. But opportunities abound in almost every sphere of education. Performance measures can be used by science teachers to evaluate the proficiency of students in doing laboratory work; by teachers of foreign languages to assess the speaking and listening capabilities of students; by music teachers to

determine the skills of students in playing an instrument or in their capacity to recognize themes and other characteristics of compositions; by English teachers to assess the public speaking skills of students; by social studies teachers to find out if students can recognize propaganda in newspaper reports. These are but a few of the many possibilities for using performance measures to evaluate instructional outcomes.

For many years the belief was widely held that there was a high relationship between scores on a written test and scores on performance measures and that the former could serve as an individual measure of the latter. This idea may have gained currency in an era when research showed a generally high correlation between written tests and final course grades. Such grades were accepted uncritically as valid indicators of overall proficiency. What was often overlooked, however, was the fact that the course grades were themselves often based on written tests, thus it was not surprising to find written tests of trade knowledge showing a high relationship with such grades. However, when special attention was given to assessing shop performance and when such performance was given appropriate weighting in the final grade, the relationship tended to go down substantially.

Today it is generally conceded that traditional written tests usually measure cognitive knowledge and are not a very dependable way to evaluate performance. Without some type of direct or indirect measure of actual performance it is unlikely that we can make an accurate assessment of an individual's overall competency.

WHY MEASURE PERFORMANCE?

Measuring performance is nothing new. For centuries it has been done informally by those responsible for transmitting skills to others --by teachers in schools, by craftsmen responsible for the training of apprentices, and by foremen in industry. At times the "test" has taken the form of a written exercise, such as a set of arithmetic problems or a spelling quiz. More often, however, it has involved the actual demonstration of a skill in a realistic setting, usually on the actual equipment of the trade involved.

Most of us recognize that there is a fundamental difference between knowing about a job and being able to do the job. "Knowledge of" is really an essential ingredient for doing a complex job correctly, but while it is a necessary condition, it is rarely a sufficient condition for satisfactory performance. During World War II, the U.S. Office of Education hired a high school English teacher to edit material relating to welding. He soon became the greatest "paper welder" in the world. He could talk a "great game" and would probably have made an impressive score on any written test about welding. Unfortunately, he had never held a welding torch in his hands and he would have been at a complete loss had he been ordered to do even a simple weld.

A person may be able to bluff on a written test, but he can seldom carry off a successful deception when a realistic performance test is required. One of the great virtues of the performance test is its impressive "face validity" and credibility, because the task one must do so closely resembles the job itself. Frequently the performance

test is nothing more than a work sample--requiring doing an actual task, but outside of the normal job environment. To save time, one may require the individual to do only part of the job, but the assignment is likely to be one calling for a relatively high degree of skill.

It is often impractical to reproduce a real job situation or to provide actual equipment. However, critical job elements can be simulated in a laboratory or in a "black box." Thus an electronic technician may be required to check out circuits and to identify and repair malfunctions on a piece of simulated equipment. Some of the reality of the work setting may be sacrificed, but the critical job elements--namely the wiring of the components that are found in complex electronic equipment--are present; thus the test is readily recognized as a realistic representation of the tasks one would encounter on the job.

A paper-and-pencil test may be easier to develop and quicker to administer and score. However, it is rarely an adequate substitute for a good performance measure. Indeed, the exclusive use of paper-and-pencil tests to evaluate students in vocational programs can have an undesirable impact on learning. Some instructors unwittingly downgrade the importance of performance by basing their evaluation of shop or laboratory work on casual observations that result in nearly everyone receiving a "satisfactory" rating on this aspect of the course. They may then use written tests to measure knowledge <u>about</u> the course. Since there may be considerable variability among the students on the written test, this test will, in effect, have a greater weight in determining the final grade than it deserves. As a result, students may be motivated to study for the written test rather than to put forth special efforts to master the performance aspects of the course.

PLANNING THE PERFORMANCE TEST

Whether one sets out to develop a written test or a performance test, the first--and usually the most difficult--task is deciding what should be covered. If the course objectives have been stated in behavioral terms, the critical behaviors appropriate to a given level of training may be identified and used as a basis for developing the test specifications. Certain of the objectives may lend themselves to testing by means of a paper-and-pencil test. This approach is the most economical and most efficient. It should be used whenever it is feasible to do so. Many performance-type problems can be presented in this format. Identification of laboratory equipment, biological specimens, or geographical features are examples that readily come to mind. The results will enable the instructor to ascertain rather quickly whether or not the student has mastered these objectives of instruction.

Whether one starts with a job analysis or with behavioral objectives which were originally derived from such an analysis, one must decide which elements are crucial to success. It is from among these critical elements that one should select the tasks to be used as a measure of performance. Because performance testing is generally a slow and time-consuming process, only a few of the critical elements can be included. One must decide which ones are really crucial. In using this criterion one assumes that if an individual is able to perform the most critical tasks in a satisfactory manner, it is highly probable that he could do equally well on other tasks which are less critical. Thus, in examining plumbers for licensing, boards frequently require the candidate to join two pieces of pipe by "pouring" or

"wiping lead." Many years ago this was probably the hallmark of the highly skilled craftsman. Unfortunately, the practice of testing this skill has continued, even though "wiping lead" itself has declined in importance.

Apart from "criticality," the test developer must consider such factors as the time required to perform a given task, the type of equipment required, the ability to present the task in a uniform (standard) manner, and the ability to evaluate an individual's performance with a high degree of objectivity. Considerations such as these impose realistic constraints on the tasks one selects to be the parts of a performance test. Often compromises are in order. Instead of requiring performance of a complex task, one may decide to limit the test to one or two phases of the task, such as preparing only one slide of a biological specimen, but identifying a larger number of mounted specimens.

Once a test plan has been developed outlining the tasks to be accomplished, details must be provided regarding such matters as equipment, materials, and procedures. The amount of detail may vary considerably depending on the subject matter involved. But the fundamental purpose remains: one must specify precisely what the examinee is to do and the conditions under which he is to do it. There should be no doubt in his mind as to what is expected of him and on what basis his performance will be evaluated. The stenographer who is given a work sample of dictation knows that she will be judged in terms of the speed and accuracy as well as the overall appearance of her work. To test a machinist's ability to use a lathe, it is necessary to provide him with specifications or blueprints of the object to be fabricated as well as the raw materials with

which he will work. He also needs to know how much time he will have and what the acceptable tolerances are for the job.

The next steps, after proper procedures have been determined, are to draft the documents that will make it possible for the test to be administered by different people at different times in a standard and objective manner.

DOCUMENTS FOR A PERFORMANCE TEST

Once the test developer is satisfied with procedures, directions, equipment, and scoring methods for a performance test, he should formalize them by preparing three documents.

1. Instructions to the Test Administrator or Observer

   These instructions outline the procedures to be followed, list the equipment that is needed, point out especially hazardous aspects or emphasize safety precautions that are applicable, and tell the administrator how to set up the equipment for the exercise. This document also defines how the test is to be scored. The instructions should be sufficiently detailed so that an administrator who is competent in the area covered by the test will be able to set it up, run through the tasks himself, and then administer the test to students in a standardized way.

2. Instructions to the Examinees

   In very simple situations, directions to the examinees may be given orally. For example, to test a musician's ability, he might be given a sheet of music and asked to play the piece. However,

such informality opens the door to the introduction of elements that could create unstandardized testing conditions. An examiner might give more detailed instructions to one individual than to another; or he might inadvertently omit something that was important from his instructions. To prevent such occurrences, it is recommended that instructions be written ones; and that they either be read to the examinee or be given to him so that he may study them beforehand. The instructions should state the purpose of the test; the time limits, if there are any; the equipment to be provided; the requirements that the examinee is expected to satisfy; special safety precautions; and information about how the test will be graded. In certain situations, some of these items may not be needed, but the person preparing the directions should make a considered judgment in each instance before omitting the information.

A major benefit of having written instructions is that each examinee receives exactly the same information about what he is expected to do. This makes for a more highly standardized testing situation. It also promotes greater confidence in the examinee since he knows that he can refer to his instructions if he becomes confused or forgets what he is supposed to do. He is able to check the equipment to make sure it is all there. He will be aware of what factors will be considered when the test is scored. Furthermore, there is no chance that the examinee may claim, later on, that the observer neglected to tell him something that was important to the successful completion of the task.

A word of caution is in order. Care should be exercised in developing the instructions to avoid revealing unintended clues

as to the proper procedure. There should be no references in the instructions that suggest what the examinee should have done earlier or what results he should have obtained from certain procedures. An alert examinee may take advantage of such unintended clues and this could give him an unfair advantage over other examinees. Some of the differences in performance might then be attributable to "test-wiseness" or reading ability rather than to the ability to perform a given task.

3. Rating and Scoring Form

A rating form should be developed for each task. This form should be a highly individualized one which specifies the checkpoints on which the individual is to be evaluated. The determination of these checkpoints is, of course, vital. There should be as many as necessary to ensure comprehensive coverage and provide reliability. Too few will probably indicate that some elements have been glossed over. On the other hand, too many elements may suggest "nitpicking" and a failure to differentiate between things that are critical and things that are trivial. The use of too many checkpoints may impose an impossible burden on the rater because he may have to watch for too many things at one time; and he may miss the important factors while trying to grade performance on minor matters. For this reason, it is urged that in developing the rating form the test developer be selective and critical. He should pick the items that are significant to successful performance. The items must be of such a nature that they can be observed and judged with a high degree of objectivity.

If experience shows that nearly everyone performs certain steps correctly, it may be advisable to omit that step as a checkpoint since it fails to differentiate among the examinees. On the other hand, if a step is important from a safety standpoint, it should be included even if the majority of candidates do it correctly. For example, if safety glasses must be worn while performing an operation, it would be advisable to include a checkpoint such as the following:

"Student wearing safety glasses? Yes___ No___
Do not allow student to proceed with
test until he puts on his safety glasses."

At certain points the observer may have to check more than one item. If a voltmeter is used in a physics project, it may be appropriate to check that it is properly connected to the unit and that the examinee has read the meter correctly. However, in many situations, it may be desirable to have the examinee record dial settings and meter readings on a separate form that is specifically keyed to his instructions.

The effectiveness of the measurement process will be reduced substantially if the observer must make judgments about quality along some sort of continuum. Experience has shown that rating scales do not work too well in performance test situations. It is preferable to design the rating at each checkpoint on an all-or-nothing basis. The examinee did or did not do what he was to do, or the measurement was or was not correct within stated limits.

* * * * *

After the various documents for a performance test have been developed, the whole package should be field tested. Tryout subjects should be proficient in the field covered by the test. They may be other instructors, practitioners of the trade, or advanced students in the subject. It is important that the tryout subjects go through all the steps so that the observer may assure himself not only that the directions are clear and unambiguous, but also that he is able to make the judgment called for at each checkpoint on the rating form. The tryout should "debug" the test and reveal any inconsistencies, errors, or even impossibilities. It should verify that all of the tasks are performable; that inserted malfunctions operate and give the indications intended; that adequate tools, instruments, and materials have been specified; and that the time limits are reasonable.

GRADING THE PERFORMANCE TEST

Whatever the nature of a performance task, sooner or later the problem of evaluating performance must be faced. This is not always an easy matter, because where one places his emphasis can be a matter of critical importance. One can focus on the product, on the process, or on both.

In attempting to arrive at a decision regarding the question of evaluating performance, a Navy chief machinist's mate who was working on a test for structural mechanics said, "I don't care if a man stands on his head while doing a job, as long as it's O.K. when he's finished." This man might be described as "product-oriented." He was concerned only with the precision attained in the final product and with its correct

operation. While most evaluators would agree that "quality of the final product" is of great importance, they are likely to argue that some consideration should be given to the "process" by which the final product is obtained. They would evaluate the individual's care of the equipment he uses, his observance of safety rules, and his adherence to approved methods of work. They might also take into account the amount of material he wastes and the time he takes to do the job.

In practice the relative weights to be given to "process" factors and to the "end product" will depend on the objectives of a particular test and the nature of the task involved. Evaluating "process" is a time-consuming and expensive procedure. Great care must be exercised in developing rating forms and in training observers. At best, results may not be as dependable as one would like because of subjective factors beyond the evaluator's control. Questions naturally arise as to how much importance should be attached to "process" ratings. In the original planning for a metals shop test, the evaluator had given equal weight to "process observations" and to "final product ratings." The instructor objected. He pointed out that a student might appear to do all the right things ("the process") and yet end up with an unusable product. He insisted that substantially greater weight be assigned to "product ratings" than to "process ratings."

In all likelihood this battle must be fought anew each time a performance test is developed, for it involves a value judgment that can be made only by those responsible for program design. There is justifiable concern, for example, that correct procedures and safety considerations--stressed in the instructional program--will be undermined if "process" is ignored

by the evaluators. If evaluation is perceived as an integral part of instruction, then this viewpoint is certainly defensible. If, on the other hand, the purpose of evaluation is to predict subsequent on-the-job performance, the major consideration should be how much the "process" score contributes to the overall validity of the performance test. Unless higher validity can be demonstrated, it is questionable that the effort and expense involved in measuring it can be justified.

"Product evaluation" is easier to deal with than "process evaluation." For one thing, the product is usually a tangible object, more durable than the fleeting actions which make up a process. Such a product may be judged without time pressure after the testing has been completed. Process evaluation, on the other hand, must generally be done while the testing is in progress.

In the case of a tangible product, it is generally easier to obtain reliable judgments regarding quality than in the case of a process. If the product is one that has been made to precise specifications (such as those given in a blueprint), it is possible to check how closely the product conforms to the specifications. An example of delayed scoring of a product is foreign language speaking tests in which the students record their spoken responses to questions, describe pictures, and read aloud. The recording may then be evaluated later for such elements as sentence structure, word choice, fluency, pronunciation, and intonation.

In situations where quality must be judged subjectively, it is important to list those characteristics which differentiate the good from the poor product and to devise techniques for measuring or otherwise assessing these characteristics. It is sometimes possible to increase

the reliability of judgments by developing a comparative scale. This may be done by having a group of highly competent judges place a number of "products" in rank order on the characteristic being rated. When a stable scale has been created (to provide benchmarks for differing degrees of goodness), it may then be used by less qualified judges to ascertain where along the scale a given "product" fits.

When the end "product" is actually a service--such as in the repair of an automobile or a television set--the judgment is generally in terms of utility. Does it work? How well? However, it is unlikely that one would be satisfied to know merely that an examinee made the repair. Part of the evaluation would hinge on how long the repair took and whether the solution was the most efficient one for the particular situation involved. Such questions inevitably take one over into the area of process, for one is now concerned with how the job was done, not merely with the end result.

Performance tests are powerful educational aids, since they emphasize doing rather than merely "knowing about." In this way they reinforce instruction because the student knows that his competency will be judged by his ability to use his knowledge and skill in a way that clearly demonstrates whether or not he has achieved the goals of the course. Performance tests are generally more acceptable to job applicants--especially those with limited verbal skills--because such applicants feel they have had a fair opportunity to show what they can do rather than demonstrate their ability to read and respond to written material.

There is no doubt that performance measures take more time to develop, that they require more time to administer, and that they are sometimes cumbersome to score. However, in terms of educational impact, it is hard to imagine a more effective approach.