DOCUMENT RESUME

ED 353 287 TM 019 025

AUTHOR	Finch, F. L.; Dost, Marcia A.
TITLE	Toward an Operational Definition of Educational Performance Assessments.
PUB DATE	92
NOTE	16p.; Paper presented at the Assessment Conference of the Education Commission of the States/Colorado Department of Educational Assessment (Boulder, CO, June 4, 1992).
AVAILABLE FROM	Riverside Publishing Company, 8420 Bryn Mawr Avenue, Chicago, IL 60631.
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Classification; Comparative Analysis; Constructed Response; *Definitions; *Educational Assessment; Elementary Secondary Education; *Essay Tests; Multiple Choice Tests; *Problem Solving; Simulation; *Student Evaluation; *Test Format
IDENTIFIERS	Alternatives to Standardized Testing; Paper and Pencil Tests; *Performance Based Evaluation

ABSTRACT

Many state and local entities are developing and using performance assessment programs. Because these initiatives are so diverse, it is very difficult to understand what they are doing, or to compare them in any meaningful way. Multiple-choice tests are contrasted with performance assessments, and preliminary classifications are suggested to promote understanding and provide an operational definition. The six suggested types of performance assessments are: (1) two-step problem solving with student constructed responses; (2) short, dichotomously scored answers provided by students; (3) short answers, essays, and "thought experiments" in which the nature of the response is up to students; (4) paper-and-pencil simulations that realistically mimic the actual environment; (5) simulations in realistic environments; and (6) evaluation in the actual environment. Five figures illustrate the discussion and provide sample problems. Six references are listed. (SLD)

Toward an Operational Definition of Educational Performance Assessments

by F. L. Finch and Marcia A. Dost

Paper presented at the ECS/CDE Assessment Conference
Boulder, Colorado

June 4, 1992

We would like to thank Jane Armstrong, Director of Policy Studies for the Education Commission of the States for asking us to discuss lessons that contractors have learned about innovative state assessment programs. We all have learned quite a bit about scheduling and budgeting as we have explored the unknown territory of educational performance assessments.

Some of the proponents of performance assessments who do not have to attend to practical considerations are much like the wiley old fur trapper who told his anxious young apprentice that this winter's project would be to locate, trap, and skin as many bears as possible. The grizzled veteran assured his apprentice that it would be easy and that he would handle the difficult tasks. The next morning the apprentice awoke to the sound of yelling and screaming outside the cabin and turned to find the old trapper's bed empty. He went to the door and looked out. He saw the old trapper running across the clearing with a bear hot on his heels. The old trapper screamed, "Open the door! Open the door!" The young trapper opened the door and the old mountain man turned aside at the last possible moment, allowing the bear to run into the cabin. The old mountain man slammed the door, dusted off his hands, and walked back toward the woods saying, "You skin this one, I will go get another one."

We all enjoy doing what we do and welcome educational performance assessments because such new ventures provide adventure, excitement, and at times even danger. Like the fur trappers, we would not be in these mountains if we did not believe there was something to be gained from charting unexplored territory.

But we also need good maps and reliable information so that we do not end up trapped in the snows like the Donner party.

Our purpose today is to share with you some of the things we have learned about student performance assessment and, perhaps more importantly, to point out some significant gaps in our knowledge. We have often compared multiple-choice tests with student performance assessments, and we will continue that tradition today because it will set the stage for what we consider to be a key issue to be resolved.

COMPARISONS BETWEEN "GENERIC" PERFORMANCE AND MULTIPLE-CHOICE ASSESSMENTS

There are so many varieties of multiple-choice tests and performance-based tests that some attempts to contrast them end up relating a subset of one form of assessment with a different subset of another form of assessment. For example, Aschbacher (1991) describes alternative or performance-based assessments as containing "tasks ... set in a real-world context or close simulation" (p. 276). Fitzpatrick and Morrison (1971) agree. However, "tasks set in a real-world context" reminds one of the many multiple-choice "life skills" tests developed in the late 1970's. It is clear that any attempt to compare multiple-choice and performance assessments should clearly state what type of multiple-choice test is being compared to what type of performance assessment.

Having said this, we can violate this recommendation by making some general observations about the two forms of assessment. For any given characteristic, it is often true that one type of assessment may be toward one end of the spectrum and the other may be toward the opposite end of the spectrum. But the two forms of assessment should usually be distinguished in terms of relative emphasis rather than by an either/or dichotomy.

The list of general tendencies on the next page was informally collected as we worked on several performance-based assessment projects and became familiar with others. We believe more strongly in some of these tendencies than others, and some are included on the basis of recommendations by colleagues.

We will make a few comments about each of the pairs without a great deal of elaboration.

It has been suggested that multiple-choice tests tend to measure a greater breadth of achievement but that performance-based tests are better able to assess the quality of what has been learned.

Some have suggested that multiple-choice tests are based on discrete skills (behavioral objectives) and performance-based assessments tend to measure broader educational outcomes. Given that performance-based assessments measure a more limited sample of the domain, it is hoped that this is correct.

A colleague suggested that performance-based tests are less contrived than multiple-choice tests. Since multiple-choice tests are more tightly controlled (each item typically measures only one factor), they tend to more closely represent scientific experiments in which the variables of interest are carefully controlled. On the other hand, especially when holistic scoring is used, it is often difficult to know exactly what is represented by the score of a performance assessment.

We believe that there is little argument that multiple-choice tests are easier to administer and to score.

In general, multiple-choice tests measure a product because we neither know, nor much care, how the student decided to select a specific answer choice. In contrast, we believe that performance assessments *should* allow us to know, or at least make inferences about, the processes used by students to arrive at a solution. Unfortunately, this does not always happen.

We believe that there is no debate about the relative cost of each type of assessment.

A critical distinction is that students select responses to multiple-choice tests and construct responses for performance-based tests.

We have often said that one of the most interesting characteristics of performance-based assessment is that it accommodates the divergent thinker. This poses some interesting challenges for those who must score the assessments. We will stipulate that multiple-choice tests tend to promote conformity in that (especially in poorly written tests) students are sometimes expected to select the answer which the item writer assumes they should select.

The format used for multiple-choice tests is certainly familiar to all. The charge that they are, therefore, boring may (or may not) be correct. The notion that performance-based tests are more "engaging" to students because they use novel procedures can sometimes be true.

GENERAL TENDENCIES

MULTIPLE-CHOICE TESTS		PERFORMANCE-BASED TESTS
Breadth of Achievement (How much?)	↔	Quality of Learning (How well?)
Skills Based (objectives)	↔	Outcomes Based (goals)
Broader Sample of Domain	↔	Limited Sample of Domain
More Controlled	↔	Less Contrived
Easy to Administer/Score	↔	Difficult to Administer/Score
Product Measures	↔	Process Measures
Costs Less	↔	Costs More
Select Response	↔	Construct Response
Promotes Conformity	↔	Accommodates Divergent Thinking
Familiar (dull?) Formats	↔	Engaging (novel) Procedures
Single-Step Problems	↔	Multistep Problems
Low Credibility Among Teachers	↔	High Credibility Among Teachers
High Credibility Among Psychometricians	↔	Low Credibility Among Psychometricians
"Scientific" Tradition	↔	Emerging Research Base
Well-Understood Characteristics	↔	Ambiguously Defined
Can Measure "Life Skills"	↔	Can Measure "Life Skills"
Does Not Simulate Instruction	↔	Mimics Instruction

That multiple-choice tests are made up only of single-step problems is often, but not necessarily, true. Performance-based assessments provide many opportunities for posing multistep problems, but developers may not take full advantage of this charaِcteristic. Enhanced multiple-choice items (discussed below) appear to require multistep problem solving.

One pair of comparisons has to do with credibility. Multiple-choice tests tend to have low credibility among teachers and high credibility among psychometricians. The reverse is true for performance-based assessments.

There is a long "scientific" tradition supporting multiple-choice tests, and it is hoped that a similar tradition will emerge in support of performance-based tests. At the present time, however, performance-based tests are supported primarily by "faith validity."

Everyone understands the characteristics of multiple-choice tests but, as we will see later, performance-based tests are, at present, ambiguously defined.

Both types of assessments can measure life skills.

Finally, one of the most heavily publicized distinctions between the two forms of assessment is that performance-based tests are designed to be patterned after, and provide a model for, teaching practices. It is not clear which is the chicken and which is the egg. Multiple-choice tests are, and always have been, designed to efficiently measure what students know and can do. It is indeed unfortunate that many educators have elected to base instruction on the format of multiple-choice tests, which are designed to be a data-gathering methodology, not a model for instruction. The lack of direct relationship between testing formats and classroom practice need not be a problem. Schoenfeld (1989) states, "If students can only employ a procedure blindly, or can only use a technique in circumstances precisely like those in which they have been taught, then schooling has in large part failed them" (p. 85).

Comparisons between multiple-choice tests and performance-based tests assume that there is such a thing as a generic multiple-choice test and a generic performance test, but our experience suggests that this is not so. The descriptions of general tendencies given above are more or less valid, but they place the two forms of assessment along a continuum in which, in some cases, they will be fairly close. In other cases, they will be fairly far apart.

It may now be useful to turn to an attempt to distinguish various types of performance-based assessments rather than to continue to describe performance-based assessments in terms of how they differ from multiple-choice tests. We suggest that various types of performance-based assessments are as different from each other as they are from multiple-choice tests.

PROPOSED CLASSIFICATIONS FOR PERFORMANCE-BASED ASSESSMENTS

We have spent so much time comparing and contrasting traditional assessments with "authentic" performance assessments that we have not carefully attended to the need to develop a taxonomy for performance assessments. A recent issue of *Applied Measurement in Education* was devoted to performance assessment. We strongly recommend that you obtain a copy of this publication, which includes an excellent paper by Pamela Aschbacher (1991) describing a state survey on alternative assessment by CRESST. One section of this article is titled "Which states are involved in performance assessment and what are they doing?" Having read elsewhere that almost everybody is creating performance assessments, we were surprised to learn that, when writing tests are excluded, only about 14 states are actively engaged in the development or use of performance assessments.

We have taken the liberty of plotting Aschbacher's data on a map. Figure 1 on the next page shows that six states are developing performance assessments and eight states have programs in place. The map also shows that nine states have some interest in this topic. At least one state, Kentucky, has moved from Aschbacher's classification of exploring possibilities to the development stage.

While this is very informative, it fails to satisfy the second part of the question -- "What are they doing?" The CRESST survey asked the state testing officers about their activities in a general way; it defined performance assessments as including "direct writing assessments, open-ended questions, hands-on experiments, performances or exhibits, portfolios of work, and so forth" (p. 277). The result is that we know, for example, that Delaware has performance assessments in physical education and geography but we still do not know, in any meaningful way, what they are really doing.

Professors teaching introductory measurement courses strongly emphasize the need to develop operational definitions as a starting point in any scientific inquiry. They suggest that an operational definition allows the researcher to describe something in a way that will allow another researcher to recreate it. It is an important goal of this paper to provide a starting point for the development of a taxonomy for performance assessments which will, it is hoped, reduce the level of ambiguity associated with this topic.

How can we communicate with each other if one of us discusses performance assessment with an assumption that it requires problem solving skills in situations which do not have just one right answer while the other person assumes that it requires hands-on activity of students? This ambiguity creates difficulties for researchers such as William Mehrens (1992), who states, "Typically what users of the term mean is that the assessment will require the examinee to construct an original response. Some people seem to call short-answer questions or fill-in-the-blank questions performance assessments. However, it is more common in performance assessment for the examiner to observe the process of the construction" (p. 3).

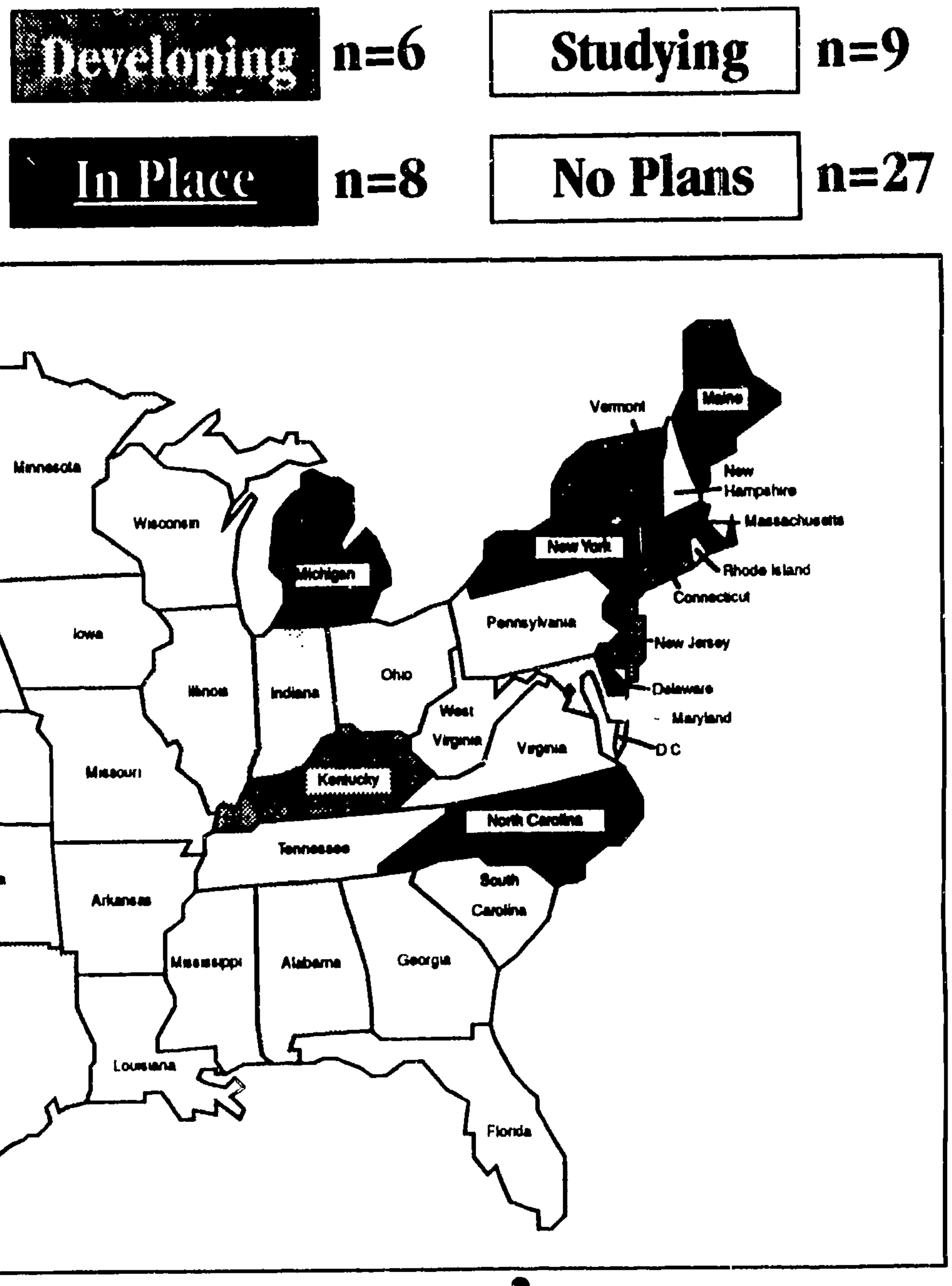
We are convinced that we must have, and can develop, a taxonomy for performance assessments which will allow us to answer the question "What are they doing?"

When an educator says "We want performance tests" and a test developer answers "We have them," it is likely that neither party has a clear understanding of what is meant by the other. This problem in semantics motivated us to enlist the aid of friends and publish a book (Finch, 1991) which we hoped would define essential terms and concepts so that we would all derive the benefits of a common vocabulary. But, as the interest in performance assessment escalated, and the number of professional papers increased, various definitions of performance assessments took off in many dimensions.

This paper represents an attempt to classify the essential characteristics of student performance assessments in terms of testing conditions and on the basis of what the students are asked to do. Figure 2 represents a preliminary taxonomy for various types of performance assessments. It is offered as a starting point for discussion and an invitation for constructive criticism.

Figure 1: State-level performance assessment activities.

State-Level Performance Assessment Activities



Source: 1990 CRESST Survey
(Aschbacher, 1991)

Riverside Chart
P.F 6-4-92

We would be especially interested in working with Pamela Aschbacher to replicate her CRESST survey using these classifications. It would be helpful to be able to describe performance assessments in terms of these classifications, but most performance assessments will be conglomerates of several types. Our experience with the California Assessment Program, for example, suggests that the proportion of performance assessment types in a specific assessment will vary not only from program to program but from subject to subject. This would require that assessments be described in terms of proportion of scoring units (not items) which can be attributed to each type of assessment.

Figure 2:

Performance assessment types.

Type 1 ---- Two-step problem solving (e.g. California's "Enhanced Multiple-Choice" items) which requires the student to construct a response and then select an answer from a list of choices (Pandey, 1991). Alternatively, the student could select a choice and then justify the answer.

Type 2 ---- Similar to multiple-choice tests except that the student provides a short, dichotomously-scored answer instead of selecting an answer choice. (Examples: 3 x 6 = _____. The main character of the story is ________ .) Usually only one possible answer, which is determined by single-step reasoning.

Type 3 ---- Short answers, essays, and "thought experiments" in which the performance requirements are clearly stated but the nature of the response is completely up to the student. Allows a wide variety of correct responses. (Includes writing explanations, constructing graphs, etc.)

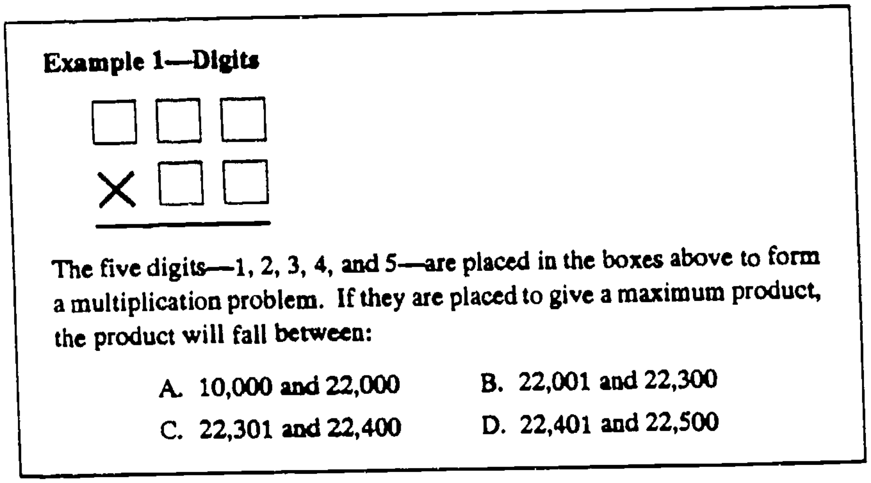
Type 4 ---- Paper-and-pencil simulations which realistically mimic the actual environment. Dials, gauges, tools, equipment, etc. are realistically illustrated. The simulation provides a major source of information for constructing a response. (Includes synthesizing text and illustrations, making judgments on the basis of supplied stimuli, etc.)

Type 5 ---- The learner is placed in a carefully constructed and realistic environment which simulates the actual situation. The task is evaluated by an external observer. (Examples: aircraft simulators and assessment "work stations")

Type 6 ---- The learner is evaluated while performing in the actual situation. (Examples: A pilot is rated by an observer while flying from airport to airport. A student conducts experiments in a laboratory to determine the chemical composition of an unknown substance.)

The numbering of the types is not intended to imply a hierarchy, with type 6 somehow being more "authentic" than type 1, but the types do provide a rough scale along a continuum which Fitzpatrick and Morrison (1971) call "fidelity of simulation" (p. 239). Except for type 1 assessments, which are based on Pandey's concept of enhanced multiple-choice items, and, possibly, type 4, it is assumed that performance assessments require students to construct a response or otherwise demonstrate competency without the benefit of answer choices. Figure 3 provides an example of an enhanced multiple-choice item reproduced from *A Sampler of Mathematics Assessment* by the California Department of Education (Pandey, 1991).

Figure 3: Example of an enhanced multiple-choice item.
(From *A Sampler of Mathematics Assessment*.
Copyright © 1991 by the California Department of Education.
Reproduced by permission. All rights reserved.)



Example 1—Digits

□ □ □
× □ □

The five digits—1, 2, 3, 4, and 5—are placed in the boxes above to form a multiplication problem. If they are placed to give a maximum product, the product will fall between:

A. 10,000 and 22,000　　B. 22,001 and 22,300
C. 22,301 and 22,400　　D. 22,401 and 22,500

Type 2 assessments are mainly characterized by short responses provided by the student; they usually have one correct answer or so few possible alternatives that they are easily scored with an answer key.

Type 3 assessments usually require a scoring rubric because they allow a wide range of possible responses. This type of assessment also requires that scorers be empowered to determine the value of unanticipated responses because this type of test provides wonderful opportunities for divergent thinkers.

Type 4 assessments represent an attempt to create, within the pages of a test book, the environment in which the task is normally performed. These assessments are usually profusely illustrated to present information as it is found in a natural environment. This type of assessment can include illustrations that range from fairly simple and partial to very realistic and complete. For example, the developer might present a pilot-in-training an illustration showing the instrument panel of an airplane and ask questions about the overall situation and specific conditions. The person being evaluated could respond in various ways ranging from answering multiple-choice items to writing a narrative solution to the problem posed. In-basket tests are also included in this type.

Types 5 and 6 represent a major shift in realism. These two types seldom use printed documents to provide stimulus materials. However, documents may be used for the examinee's response and, more often, used for the evaluation record created by a observer who scores or rates the examinee.

A type 5 assessment consists of an artificial simulation which has been constructed to carefully simulate the target environment. An example might be an "evaluation station" which contains the materials required for the examinee to demonstrate the capability of performing a specific task. For example, a science teacher might set up a testing station to determine whether a student knows how to identify acids and bases. In its simplest form, the materials available might be litmus paper and several bottles containing unknown liquids. The assessor could make the task more difficult by including materials which are not needed to perform the task.

Type 5 assessments may also be developed by using realistic computer graphics which include both visual and auditory stimuli.

Type 6 performance assessments place students in the actual setting and give them complete freedom to use whatever materials are available to perform the task to be evaluated. For example, the science teacher might bring a student into a fully equipped chemistry lab and ask her to use mercuric oxide (mercury II oxide) to create liquid mercury and oxygen gas. In the real world, pilots are periodically evaluated by an observer who merely observes the pilot conduct the pre-flight check, get the plane off the ground, do whatever must be done to fly from one point to the other, and land safely. The Air Force has a very simple criterion to evaluate the competence of pilots: Takeoffs equal landings.

"Decision clues" associated with the six types of assessments are presented on the next page. They can provide additional guidance for classifying various types of assessments. Like the proposed types described above, this information may be considered a work in progress. Comments and suggested revisions will be welcomed.

Figure 4:

Characteristics of performance assessment types.

Decision Clues	Types					
	1	2	3	4	5	6
A. Guidance about the nature of the desired response	Obvious	Obvious	Much	Little	Varies(2)	None
B. Paper-and-pencil test response format?	Yes	Yes	Yes	Often	Seldom	Never
C. Range of variability in responding	Little	Some	Great	Varies(1)	Great	Great
D. Freedom of response choice	Little	Much	Much	Varies(1)	Great	Complete
E. Description of the goal	Obvious	Obvious	Specific	Varies(1)	Varies(2)	General
F. Scoring/Evaluating	Product only	Product only	Usually Product	Varies(1)	Process(3)	Process(3)
G. Evidence about process used	None	None	Some	Varies(1)	Much	Much

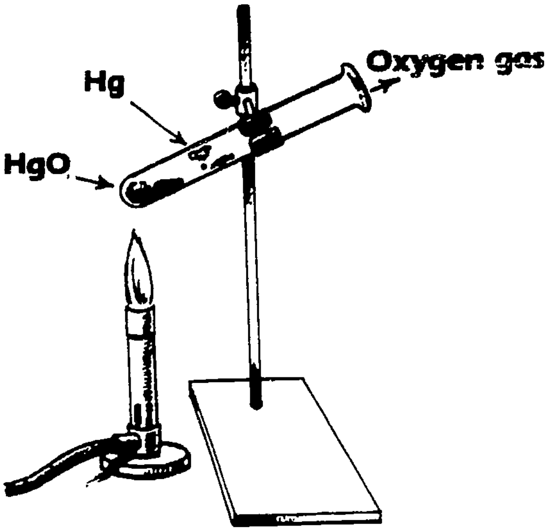
(1) Could be multiple-choice, essay, or observed performance

(2) Depends upon the way the task is presented

(3) Includes "performances" and products which result from exemplary application of process skills (e.g. Sistine Chapel)

A sample science test is presented in Figure 5 to show a variety of exercises which may be classified according to the first four types discussed earlier. It is perhaps realistic to assume that paper-and-pencil performance assessments will typically contain many types of exercises. The reader is invited to classify the six elements of this test according to the information presented earlier.

Figure 5: Sample science test.



Description of the Experiment

1. A red powder, mercury (II) oxide, is placed in a test tube and heated. The powder turns dark. Silver-colored drops begin to form on the sides of the test tube and oxygen gas is given off. What has happened? Choose one or more answers.

   A. A compound has been broken down.
   B. Elements have been created.
   C. Oxidation has taken place.
   D. Decomposition has taken place.

2. What will probably happen if some of the dark powder is allowed to cool and oxidize?

   ____

   ____

3. When allowed to cool, some of the dark residue remaining at the bottom of the test tube oxidizes and turns back into mercury (II) oxide. What color will it be? ______________

4. What is the chemical symbol for mercury? ______________________________

   The balanced equation for this reaction is:

   $$2Hg0(s) \longrightarrow 2Hg(1) + 0_2(g)$$

5. In the equation, what is (g)?

   A. a mathematical symbol
   B. the weight of the atom
   C. one of the three states of matter
   D. an indication that the reaction has reached chemical equilibrium

6. How many atoms of oxygen are on each side of the equation? __________

Exercise 1 is a traditional multiple-choice item. The fact that the students can choose one or more answers does not suggest that this item is either an innovative item or an enhanced multiple-choice item. By the way, the answers are A and D.

Exercise 2 is a type 3 problem because it requires a fairly elaborate response and has multiple correct responses. In constructing a scoring key we would give one point for a response which states, in effect, "It will become red again." We would also give one point if the student responds "It will absorb oxygen." We would give two points if the student mentioned both of these factors in isolation and would give three points if the student linked them by saying that "It would absorb oxygen and turn red because it becomes mercuric oxide." We would give no points if the student said "It will oxidize" because this is given in the statement of the problem.

Exercise 3 is a type 2 problem because, although it allows a free response, there is only one possible answer -- "It will become red."

Exercise 4 can be misleading because it requires only a short answer. The correct response is Hg. This is a type 4 assessment because, while the response is simple, the student must synthesize a great deal of information to reach this conclusion. In exercise 1 it is stated that the red powder is mercury (II) oxide, but the student must combine this statement with symbols provided in the illustration of the experiment to answer the question. The illustration shows the symbol Hg, but the student is not told that this is mercury. The stem for exercise 1 deliberately refers to it as "silver-colored drops" to avoid a simple matching of information. The student can also answer this question by referring to the balanced equation between exercises 4 and 5 and determining that Hg and O_2 are the result of the experiment but must relate this to other information to associate Hg and mercury. There are many ways for a student to provide a simple answer to this question, but all of them require a multistep analysis of available information.

Exercise 5 is an enhanced multiple-choice item because students must solve a problem which is not stated before they can select a correct response from the information provided. One can reasonably infer that (s), (l), and (g) stand for solid, liquid, and gas. If the correct answer were "gas," this would be rather ordinary multiple-choice item, but the choices require that the student make inferences and then link those inferences to the knowledge that the three states of matter are solids, liquids, and gases.

Exercise 6 is obviously a type 2 assessment.

Summary and Conclusions

Many states and local entities are exploring, implementing, and developing educational performance assessment programs. These initiatives are so diverse that it is very difficult to understand what they are doing or to compare them in any meaningful way. It is hoped that the preliminary classifications suggested in this paper will promote understanding. It has been said that unless one's information is well organized, the more of it you have, the less you will know.

REFERENCES

Aschbacher, P.R. (1991). Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education, 4*(4), 275-288.

Finch, F.L. (1991). *Educational performance assessment*. Chicago: The Riverside Publishing Company.

Fitzpatrick, R. & Morrison, E.J. (1971). Performance and product evaluation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.) (pp. 237-270). Washington, DC: American Council on Education.

Mehrens, W.A. (1992). Using performance assessment for accountability purposes. *Educational Measurement: Issues and Practice, 11*(1), 3-9, 20.

Pandey, T. (1991). *A sampler of mathematics assessment*. Sacramento: California Department of Education.

Schoenfeld, A.H. (1989). Teaching mathematical thinking and problem solving. In L.B. Resnick and L.E. Klopfer (Eds.), *Toward the thinking curriculum: Current cognitive research* (pp. 83-103). Alexandria, VA: Association for Supervision and Curriculum Development.