Testing for Learning Outcomes

TEST! QUIZ! EXAM!
PERFORMANCE APPRAISAL!
EVALUATION!

Do these words make your palms sweat? Increase your heart rate? Keep you awake at night? If so, don't feel alone. Dread of being evaluated is a feeling shared by grade school students and company executives alike.

Of course, the company line is that it's for your own good.

And it's true—provided that the evaluation is done right and results are used constructively. Unfortunately, most of us have had a bad experience somewhere along the way that has made us nervous about tests. Here are just a few typical testing and evaluation nightmares that are responsible for our test phobias:

- After listening to their instructor for an entire course, learners still don't have the slightest idea what they'll be tested on.

- Learners work hard to master complex skills and ideas covered in a particular course. But evaluation consists of multiple choice and true-false questions that test their memories for minor details.

- Test questions are written so badly that learners can't figure out what's being asked.

- The instructor, annoyed with restless learners, punishes the class by giving them a pop quiz they're not prepared to take.

- Learners encounter trick questions designed to make sure no one gets a perfect score. (Too many As make the course seem too easy.)

- An employee evaluation is based on an unrealistic job description; the employee is evaluated on skills irrelevant to the job, while skills critical to the job are ignored.

If you haven't experienced any of these situations, you are a lucky person and probably have no trouble sleeping the night before you are evaluated.

Before you get complacent, however, consider that while you may have no problem, you may inadvertently be the cause of someone else's nightmares. If you've ever written a test, chances are you've written a bad test question. The best way to eradicate test phobia is for evaluators themselves to try to understand and adopt good evaluation practices.
Types of Tests

Let's look at some of the main testing methods. You've probably experienced most of these methods, but one or two are a bit exotic, not often seen outside special training situations. Written test items can be divided into two major groups:

- objective items, and
- subjective items.

OBJECTIVE TESTS

Objective test items are those that have only one correct answer. They include:

- multiple choice,
- matching columns,
- true/false, and
- fill-in or completion items (when the correct answer is only a word-or its synonym-or phrase).

They are called 'objective" items because they can be objectively graded. Anyone with an answer key can grade them with identical results, and the grader doesn't need to make any decisions. Objective questions are used in computerized testing because the correct answers can be easily programmed.

Advantages and Disadvantages of Objective Tests

Time and effort: Good objective test items are difficult to write and take a lot of thought. But once they are developed, they can be administered by one person, and are easy to correct. They can even be corrected by a machine or by an assistant unfamiliar with the subject matter.

Cost: The main expense of using objective tests is in the initial development. But once developed, the test items can be refused with minimal cost.

While objective tests are economical (after they've been developed), they are sometimes misused because they aren't always the appropriate means for testing learning objectives.

SUBJECTIVE TESTS

Subjective test items have many possible answers. Examples of subjective items include:

- essay questions (usually several paragraphs or pages), and
- short answer items (requiring a short explanation of a few sentences).
An instructor or subject matter expert must grade each item and decide whether or not it meets the criteria of acceptability.

**Advantages/Disadvantages of Subjective Tests**

Time and effort: Subjective test items are easier to write than objective items, but the developer should take care that the questions are clearly written and well-organized. The decision-making process of grading is time-consuming, as is writing a detailed answer key to ensure that tests are graded using the same criteria.

Cost: Cost of development is generally low, but grading costs are higher since only instructors or subject matter specialists can grade them.

**ORAL TESTS**

Oral test items are usually subjective. Types of oral testing in training include:

- panels of subject matter experts who gauge the depths of the learner's knowledge
- walk-throughs, in which the learner walks through a task and explains a procedure and/or points out locations of components in a plant or on equipment
- talk-throughs, in which the learner explains, step by step, how a particular task is carried out.
WRITING GOOD OBJECTIVE TEST ITEMS

TRUE/FALSE TEST ITEMS

True/false test items often are used because they are easy to write, easy to correct, and don’t take up a lot of class time.

True/false testing should be used cautiously, however the method can be invalid and unreliable because:

- questions are easy to guess (there's a 50% chance of being correct)
- knowledgeable learners tend to be tricked by them because they are aware of exceptions.

True/false questions should be used only to test knowledge objectives. Technically speaking, they are valid only for objectives that ask learners to recognize or identify correct information rather than to state or recall it.

Guidelines for True/False Items

Test only one idea at a time. Learners may know the answer to one part of the question but not the other. Then it’s impossible to know what they know and what they do not know. More than one statement in a question also may be confusing.

Example: Thomas Edison invented the light bulb and Henry Ford invented the automobile. This is a weak true/false item because it asks about unrelated events.

Avoid ambiguous questions. Words like "seldom," "often," and "possible" tend to confuse.

Example: It seldom rains in Greece during the summer. How often is seldom—once a week? Once a month? Where in Greece—in the mountains? On the islands? In the north?

Avoid such words as "always," "never," and "none." Many learners realize that such definite statements usually are false, so they will guess without really reading the question. Or they may get the question right for the wrong reason.

Example: Pure water always boils at 212°F. The learner may get this item right for the wrong reason. He or she may answer "false" because of the word "always," rather than because water boils at 212°F.
MULTIPLE CHOICE ITEMS

Multiple choice questions are difficult to write. They are common because they are easy to grade and don’t use a lot of class time to administer. The problem is that often they are misused and poorly written.

Multiple choice items are most appropriate for:

- knowledge-based objectives in which the learner needs to choose correct information
- simple problem-solving skills where it is appropriate to select a certain solution.

Guidelines for Multiple Choice Items

Put the blanks toward the end of the main part of the question (the stem). This makes it easier for learners to read and understand what is being asked.

The stem is usually followed by four or five possible responses. There should be only one correct response; the rest should be distractors-words or phrases meant to distract the learner’s attention away from the correct response.

Writing good distractors is the most difficult part of constructing good multiple choice items.

All distractors should be believable. One or more obviously-wrong distractors makes the correct answer easier to guess.

*Example: The capital of Maryland is (a) Baltimore (b) Aberdeen (c) Columbia (d) Annapolis.*

Or:

*The capital of Maryland is (a) Miami (b) New York City (c) Kansas City (d) Annapolis.*

The first of these items has plausible distractors. The second item has such poor distractors that the correct answer is obvious to most people.

Distractors should agree grammatically with the stem so they don’t provide clues to the learner.

Distractors should be about the same length as the correct response. A response that is substantially longer than the rest often is the correct one.

The correct responses should be placed in random positions. Test-wise learners know that (c) and (b) are favorite spots for correct responses. It is also important not to develop a pattern of correct responses across the test.
Using Multiple Choice for Math Problems

Multiple choice items should be used to test calculation objectives only with extreme caution. Test-wise learners often can pass multiple choice math tests with a minimum knowledge because of implausible distractors.

Multiple choice items should be used to test math problems only if the distractors are very good. One way to construct good distractors is to go through the problems, making different common mistakes for each distractor. Warning: changing the decimal point is not a difficult-enough distractor to fool all learners (unless it is a common mistake for that particular problem).

MATCHING COLUMN ITEMS

Matching column items are a lot like multiple choice items, so many rules of construction are the same. The matching items have directions, the problems, and the distractors.

Example:

Directions: Match the part of a gas combustion engine to the function it performs.

<table>
<thead>
<tr>
<th>PART</th>
<th>FUNCTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>(problems)</td>
<td>(distractors)</td>
</tr>
<tr>
<td>1. Alternator</td>
<td>a. Cools the engine</td>
</tr>
<tr>
<td>2. Carburetor</td>
<td>b. Ignites fuel</td>
</tr>
<tr>
<td>3. Spare plug</td>
<td>c. Keeps track of mileage</td>
</tr>
<tr>
<td>4. Radiator</td>
<td>d. Mixes oil with gas</td>
</tr>
<tr>
<td></td>
<td>e. Mixes gas and air</td>
</tr>
<tr>
<td></td>
<td>f. Recharges the battery</td>
</tr>
</tbody>
</table>

Guidelines for Matching Column Items

All the problems should be related so the distractors sound reasonable.

Distractors should agree grammatically.

There should be more responses than problems so the last problem will not automatically match the last response.

Responses should be in a logical order such as alphabetical order, or, if numbers are used, in numerical order.

Every item should have directions. If responses can be used more than once, the directions should indicate that fact.
FILL-IN OR COMPLETION ITEMS

Fill-in items are generally easy to construct. They can be used for information recall, or for short responses to mathematical problems.

Examples:

__ A univalve is a mollusk with one ____________________
__ the area of a square with 3-inch sides is _____ square inches.

Guidelines for Fill-in Items

The context of the question must be included; the learner should not have to guess what the question is asking.

Examples: An alligator is a(n) __________.
Better: An alligator is in the phylum ________________.

In the first example, the learner can't tell how to answer the question. The instructor would be forced to accept any correct response, such as animal, lizard, reptile, etc.

The blank should come toward the end of the statement to lessen confusion.

Examples: A(n) ______________ is a hybrid from a donkey and a horse.

Better: The hybrid offspring of a donkey and a horse is a(n) ______________.

No grammatical clues should be given. Using a/an, a(n), he/she, him/her, etc. will prevent the learner from eliminating items through grammatical clues.

Advantages and Disadvantages of Oral Tests

Time and effort: Since the tests usually are administered with one or more instructors (or subject matter experts) and one learner, oral tests are time consuming.

Cost: The cost factor of the instructor’s time makes oral testing expensive.

Certain advantages exist for the learner in an oral testing situation that do not exist in a written testing environment.
For example, if the learner doesn't understand a question, the instructor can reword it. The instructor may also prompt the learner to give more information by asking further questions. Oral testing is a good way to explore the depths of a learner's knowledge because there is always an opportunity to ask more questions.

**PSYCHOMOTOR TESTS**

Psychomotor testing (also called performance testing) is the actual performance of a physical task using real equipment or a simulator. These tests allow the instructor to see if the learner has the physical and mental skills to perform given tasks.

All physical skills should be tested by a psychomotor test. Any time an objective asks the learner to do something like "measure," "adjust" "operate" or "repair," it should be tested by actual physical performance of that task.

Testing of physical skills should be done in the same or a similar environment, and under similar conditions to those on the job. If a mechanic is being tested on car repairs, he or she should be tested in a garage that has the same equipment used on the job.

Sometimes it isn't possible to test in the actual environment because of costs or dangers involved in on-the-job testing. Examples include nuclear plant operations or fire control training. In those cases, tasks may be tested in a shop, laboratory, or simulator.

An instructor or supervisor monitors and rates the learner's performance. In order to ensure that each performer is rated by the same standards, he or she uses checklists that define the standards of acceptable performance. See the sidecar on page 5.

**Advantages and Disadvantages of Psychomotor Tests**

**Time and effort:** Psychomotor tests are time-consuming because the learner and instructor are usually in a one-to-one testing environment.

**Cost:** In addition to the cost of the instructor's time, psychomotor tests can be expensive because they usually involve the use of costly equipment or simulators.

Despite the cost, a physical demonstration is the only valid way to test psychomotor skill. For example, some electrical utilities use simulators to train the personnel who buy and sell electricity on the national grid. These people must make quick decisions that will affect the cost of electricity to the consumer
CHECKLIST CONSTRUCTION FOR PSYCHOMOTOR TESTING

Here is one possible set of steps for developing checklists for psychomotor testing:

1. Analyze an expert performing the task. Determine the critical steps and key decisions that must be made by the learner in order to perform the task correctly.

2. Set the guidelines for the checklist construction. How will it be organized-through:
   - sequential steps?
   - a product measure?
   - criteria such as speed, safety, and accuracy

3. Determine the rating scales to be used. Will there be:
   - a two-point scale (such as yes/no or pass/fail)?
   - a numerical scale grading performance on a scale of 1 to 10?
   - a graphic rating scale? (See the example.)

4. Spell out guidelines for rating learner performance. Remember, there is a certain amount of subjectivity to rating the performance of physical tasks. In order to ensure that all raters will grade by the same criteria, it is necessary to write a set of guidelines and discuss them with raters before the test is administered.

GRAPHIC RATING SCALE

<table>
<thead>
<tr>
<th>Engine starts immediately.</th>
<th>Engine hesitates. There are knocks and pings.</th>
<th>Engine doesn't start.</th>
</tr>
</thead>
<tbody>
<tr>
<td>_________________________</td>
<td>_________________________________</td>
<td>_____________________</td>
</tr>
</tbody>
</table>

Checklists can be organized in the following ways:

- steps of the task in sequential order
- criteria of performance (categories such as safety, speed, technique)
- product measure (testing the final product or result for standards of quality)
- a combination of the above.
WRITING GOOD SUBJECTIVE TEST ITEMS

A subjective test should not test the learner’s ability to understand complex questions and organize essays unless that is clearly a requirement for the course. In order to avoid testing essay writing as a hidden agenda, write essay questions so that they require only a few paragraphs each. It should be made clear—both orally and in writing—that brief answers are expected.

Questions should be short and concise. If one question covers a lot of information, break it up into sections. This helps the learner to understand the question, and the instructor to organize the answer.

Look at the following examples of subjective questions for customer sales representatives at a bank. Which one is easier to understand?

Describe the procedure for selling and opening a money market account: how to determine who’s a perspective customer, advantages of a money market account over other investments, and rates of interest and paper work to be completed.

or:

Write short descriptions of the steps for selling a money market account. Include:
(a) determining the customer’s needs
(b) introducing the idea of a money market account
(c) describing the advantages of a money market account over other investments
(d) explaining interest rates and how interest accrues (a) completing the paper work for the customer

The second item organizes the learner's response and lets the learner know exactly what the instructor wants.

Steps for Writing Subjective Questions and Grading Keys

1. Write a question. Check it for clarity. Compare it with the objective to ensure a match.
2. Write the expected answer.
3. Ask a colleague to read the question, checking it for clarity.
4. Ask the colleague to write an answer.
5. Revise the question (if needed), based on answers written by the instructor and the colleague.
6. Construct a grading key. The key should include all components of a correct answer, and the point value for each piece of the answer. Make sure that point values assigned to different parts of the answer are weighted according to their importance.
Playing Fair

Every time a course designer constructs a test item, he or she should make sure the test is as fair as possible while giving learners who know the material a chance to prove what they know. The main point of evaluation is to determine what skills and knowledge a learner possesses.

KEEPING LEARNERS INFORMED

Objectives serve the purpose of informing learners what is expected of them. Each session should begin with an introduction addressing objectives to be covered during that session. This way, learners constantly know where they are in the course, and what is to be accomplished.

Objectives should also be used for review before testing. Learners should be told which objectives are being covered by the test, and how they will be tested. Informed learners are motivated and in control of their learning process. They are more likely to succeed than those who do not know what is expected of them.

WRITING CLEAR, CONCISE TEST ITEMS

It is difficult to write good test items that are easily understood. One way to ensure good test construction is to have someone else review the test to make sure it's written in plain English (or Dutch or Portuguese). The instructor can also ask the reviewer to answer the questions to see if the test yields the expected answers.

GIVING AWAY THE ANSWERS

Objective test items are notoriously easy to guess by test-wise learners. Some of the common clues found in test items are:

grammatical hints -- disagreement among the number and tenses in distractors, or use of "a" and "an" instead of "a/an" in the question

word cues -- using a key word in a question that is also in the distractor or in the answer. Example: asking what piece of equipment transforms current into voltage, and writing "transformer" as the answer
HIDDEN AGENDAS

Test items should only test the skills and knowledge they were meant to test. But sometimes test items have a hidden agenda. That means that a learner is required to do or know something outside the prerequisites and requirements of the course. For example, asking a bank teller trainee for well-organized essays on customer service would call for written communications skills unnecessary for the training program and the job.

GOOD GRADING PRACTICES

Subjective tests should be graded using the same criteria—a difficult challenge. Here are some problems and possible solutions:

Grading differences among instructors -- If a test is standard and given by more than one instructor or more than one person is grading, learners may be rated differently. A key should be used, detailing the grading criteria. All raters should meet and discuss how these criteria will be applied. Decisions to give partial credit must be uniform.

Ensure continuity by:

- having each instructor grade selected questions across the entire test.
- having one person look over all tests at the end of the grading process, and standardize marks.

Order of grading. It is difficult to rate performance by consistent criteria. If the initial performances are very good, the grader tends to judge the following ones harder. If initial performances are poor, succeeding ones tend to be judged more leniently. Reviewing all papers one more time helps to ensure equal application of criteria.

Prejudice based on past performance. When a grader rates a learner he or she knows, it is difficult to separate current performance from past performance. Good past performance tends to work favorably for the learner, while poor past performance may unfavorably color the rater's evaluation of current performance. On written tests, names should be covered or numbers assigned.

TRICK QUESTIONS

Some instructors try to fool learners by asking them trick questions. Unless an item tests for recognition of a specific tricky situation that is part of the necessary training, trick questions are simply not a fair test of a learner's skills and knowledge. Sometimes an instructor will inadvertently write a trick question. He or she must be open to removing or changing an unfair question when it is pointed out.
Evaluating Test Results

Two methods of evaluation are:

**Norm-referencing** -- a traditional and familiar means of evaluating learner performance. It's how we often were graded in school, but it's not always the best way to evaluate performance.

**Criterion-referencing** -- another perspective on evaluation.

**NORM-REFERENCED EVALUATION**

Remember back in school when the class did badly on a test but the teacher assured us that it would be "graded on the curve"? Remember the importance of class rank for students who wanted to get into certain colleges?

Class ranking and grading on the curve are artifacts of norm-referenced evaluation. Norm-referenced evaluation compares an individual's performance with that of all the others taking a particular test or in the same class. The worse the rest of the population does, the higher the individual scores.

When a large number of people take a standardized test, their marks usually fall within a normal range. This range, when plotted, forms a bell-shaped curve like the one below. It shows the performance of most people falling in the middle, with a few doing very well and a few doing very poorly. When scores are marked on a curve, marks are adjusted so that they still form a bell curve, but instead of an "A" starting at 100%, it may begin at 75% or 80%.

This seems like a strange way to evaluate whether or not a learner can perform a task or demonstrate knowledge. If the best mark is an 80%, does that mean that none of the learners were able to perform the task completely?

The truth of the matter is that norm-referencing evaluation should not be used to determine who can and who can't perform a task. The purpose of norm-referencing is to identify the best and worst performances by comparing each individual to the rest of the group. It may be used in cases like:

- There is room in a training program for five learners. The company wants to give the spots to the best five applicants.

- A scholarship will be given to the top quarter of the class. They need to be identified.
• An organization is offering remedial math and reading courses. Employees who need them most will be admitted into the courses first.

CRITERION-REFERENCED EVALUATION

Criterion-referencing means that an individual's performance is measured against the standards of the objectives being tested. This type of evaluation should be used in any program in which learners must master particular skills or knowledge before continuing. Criterion-referenced evaluation enables learners to compete against their own performance rather than others'. In order to reach their goals, they must meet the necessary objectives.

Most organizations are interested in whether or not an individual has the skills to perform a particular job. They aren't interested in the individual's ranking among 20 other trainees who may or may not have the necessary skills.

Still, instructors tend to compare learners' performances without knowing whether or not any of them have the skills and knowledge it takes to continue. This is a mistake made in both educational and training environments.

Criterion-referenced evaluation is based on meeting the objectives. The most appropriate way of grading a course that uses criterion-referenced evaluation is pass/fail. When learners pass, it means they have mastered all the skills and knowledge needed to perform specific tasks.