The Two Most Important Things You Can Do to Improve Testing in Your Organization

William Coscarelli and Sharon Shrock

William Coscarelli & Sharon Shrock

618.453.4217 618.453.4218
coscarel@siu.edu sashrock@siu.edu

Curriculum & Instruction, Southern Illinois University
Carbondale, IL 62901-4610

Bill Coscarelli and Sharon Shrock are Instructional Development faculty at Southern Illinois University. They have served as consultants to universities and businesses regarding testing and evaluation issues and recently, in a unique university and corporate relationship, co-directed the Hewlett-Packard World-Wide Test Development Center. They have provided testing guidance to over 25 global companies.

They have written on CRT issues in a number of journals. The first edition of their book, *Criterion-Referenced Test Development: Technical and Legal Guidelines for Corporate Training*, won the Outstanding Book in Instructional Design from AECT's Division for Instructional Development in 1990, and the Outstanding Instructional Communication Award from ISPI in 1991. The second edition is now in press with the International Society for Performance Improvement.

At the SITE 2000 Conference in Nashville we presented a session entitled "Testing Triage: Maximizing Effectiveness in Assessment with Minimal Investment." We used the term "triage," which means determining priorities for action in an emergency, to present seven items of advice that can be used to vastly improve the quality of testing in an organization with little additional investment of time or resources.

It has been our experience that most of the time and money organizations spend on testing is wasted because the tests are invalid, i.e., they don't measure what they are supposed to measure. Many organizations that do testing often think that sophisticated testing is too expensive. In the insurance field in particular, there are excellent models of well developed tests that rely on state of the art delivery and analysis for certification of professionals. However, for most in-house training settings the expertise of these organizations is neither available nor necessary. We believe that in most instances testing correctly is no more expensive than testing incorrectly, and that testing incorrectly is a waste of resources and creates a culture of anxiety, avoidance, and often outright hostility among trainees who are tested.

At the end of our session we talked with a number of participants who challenged us to further refine our advice. After thinking about their questions and concerns, we have triaged our triage and offer our advice on the two most important things you can do to improve testing in your organization:

1. Write test items above the memorization level.
2. Use the Angoff method to establish the cut-off score for mastery.

Bloom's Cognitive Classifications

In 1956 Benjamin Bloom and his colleagues created a taxonomy of cognitive objectives. The taxonomy was created to improve precision in the testing of cognitive

processes. The classification scheme consists of six levels with each given level subsuming all levels beneath it as follows:

- Evaluation
- Synthesis
- Analysis
- Application
- Comprehension
- Knowledge

Each of these cognitive levels is described in great detail in the book *Taxonomy of Educational Objectives*, and these levels have stood the test of time. Understanding the nature of the cognitive performance to be assessed is a good first step to being able to write an appropriate test item. In Bloom's book the description of each cognitive level is accompanied by many examples of test items that assess that particular cognitive behavior. If a test writer can correctly identify the Bloom level of an instructional objective, a wealth of ideas about how to measure the objective become available.

Another important result of understanding Bloom's Taxonomy is an increased awareness of all the cognitive behaviors beyond simply remembering, i.e., beyond the Knowledge level. Most of the tests we take in school at all grade levels and even at the college level are composed of knowledge level questions. This circumstance is not difficult to explain, since Knowledge level items are by far the easiest to write. However, developing tests that truly reflect on-the-job performance requires the ability to distinguish among different cognitive behaviors and skill in writing items at the higher cognitive levels, particularly the comprehension, application, and analysis levels.

Write Test Items Above the Memorization Level

In general, the single most useful improvement you can make in writing test items is to write them above the memorization level. As just noted the vast majority of test items are written at the memorization level. In contrast, the vast majority of jobs require performance that is above the memorization level. This disconnect between testing practice and job performance is what often leads management to question the value of training and turns testing into a misleading indicator of performance, e.g., "How come they passed the course but can't do the job?" is a common summary of the problem. When you design your test, first consider the job, and then consider the level of learning your test assesses in light of this job performance. In Bloom's terms, design your test items above the "Knowledge" level. It is usually not productive to worry about precise classification of items beyond Knowledge level; the critical distinction is between memorization and everything else above it. Let the job drive the cognitive level of the test items.

Memorization level items are usually items that ask the test taker to define a concept or rule or remember exactly what was presented in a presentation or manual. "Everything else" types of items usually ask the performer to apply the definition of a concept or rule in test items that are composed of previously unencountered examples providing the scenarios which test takers must comprehend, solve, analyze or judge.

It is essential that you distinguish between the tasks that require simple memorization items as opposed to those needing scenarios that show what the performer must face on the job. Asking an agent trainee to identify the steps of establishing rapport does not mean that they would recognize whether or not those steps where followed or be able to apply them in the office.

Here is an example of a memorization level item:

An endowment insurance policy:

(a) is death protection for a specific period of one or more years.

(b) gives death protection for as long as the policyholder is alive.

(c) pays a sum or income to the policyholder if they live to a certain age.

(d) allows the amount of insurance (and the related premiums) to be increased or decreased as needs change, subject to evidence of insurability.

Now compare the level of difficulty and applicability of that memorization level item with these two "everything else" items:

Jay is a 28 year old graduate student just beginning studies for a Ph.D. in philosophy. He has four children. His wife stays home to care for and school the children. His assistantship pays $920/month. The best type of policy for him would be:

(a) Term

(b) Whole Life

(c) Endowment

(d) Universal Life

Michael was flying his plane when he encountered a fierce windstorm just as he was running out of gas on an approach to landing. Unable to make it to the airport due to the headwind he landed on South Illinois Avenue just as the student bars were letting out. As the students saw the plane and ran for safety they trampled Dr. Post's four Bradford Pear trees valued at $1200 each. (Dr. Post has a $250 deductible on her homeowner's policy.) Dr. Post files a claim for damage to her trees.

Is she entitled to recover damages? And if so, how much?

As you can see, the "everything else" items look more like what agents or claims adjusters have to do on the job. Getting these higher level items correct demonstrates that test takers have remembered the required knowledge level material as well. Testing at the "everything else" level is the most important thing you can do to improve your tests.

Use the Angoff method to establish the cut-off score for mastery

Unfortunately most cut-off scores for tests are set the way test writers (and test takers) remember their high school teachers' grading scale—70 was passing. Setting a cut-score in such an arbitrary manner is misleading at best, hazardous at worst, and usually legally indefensible. The good news, though, is that there are a general class of techniques for estimating the cut score based on subject matter assessments of item difficulty.

This class of techniques is often referred to as Conjectural Methods in that they rely on professional estimates or conjectures of success. Of the conjectural techniques, the Angoff method (Zieky & Livingston, 1977) is perhaps the most useful and generally used and the one we want to discuss.

1. The first step is to identify judges who are familiar with the competencies covered by the test, and with the performance level of masters of these competencies. The number of judges you select will depend on availability of judges, criticality of the performance, etc. However, we think you would rarely need more than five, with three being the more typical number.

2. The judges are then asked to review each item in the test. For each item, each judge estimates the probability that a minimally competent test-taker would get the item right. Make sure the judges understand that a probability level should

never be lower than the level of chance predicted by the item, e.g., if there are four alternatives in a multiple-choice item, the estimate should not be lower than 25%.

These estimates are expressed as percentages and assigned a corresponding decimal value. For example, if a judge thinks there is a fifty-fifty chance of the minimally competent test-taker getting a given item right, that item is assigned a value of .50. If the judge estimates that an item is so simple or so critical that the minimally competent test-taker will almost surely get it right, then the item would be assigned a value of 1.0.

If possible, judges should estimate the probability for each item independently, and then discuss among themselves those items where they disagree markedly in their estimates.

3. The chosen cut-off score is the sum of the probability estimates. If more than one judge is used, the cut-off score is the average of the sums.

Here is an example (Table 1) of the process illustrated with a five item test:

Table 1

Judges' Probability Estimates

Angoff Method

Item	Judge 1 Probability	Judge 2 Probability	Judge 3 Probability
1	.33	.50	.40
2	.80	.90	1.0
3	.20	.33	.20
4	.20	.90	.33
5	.50	.75	.50
Total	2.03	3.38	2.43

Averaging the Totals for Each Judge

to Obtain the Cut-Off Score

2.03 + 3.38 + 2.43 = 7.84

7.84 / 3 = 2.61

Cut Score = 2.6

Final Comment

Good testing practices benefit both the individual and the organization. For the individual, tests can provide feedback on personal competence. For the organization they can provide information on whether or not people can do a job or task that needs to be done. As we said earlier, in most instances testing correctly is no more expensive than testing incorrectly. Increasing the knowledge and awareness of the test designer is the only required investment.

Bloom, B. S. (Ed.). (1956). *Taxonomy of educational objectives.* New York: David McKay Company, Inc.

Zieky, M., & Livingston, S. (1977). *Manual for setting standards on the basic skills assessment tests.* Princeton, NJ: Educational Testing Service.