

# Measuring Competency: Understanding the Tradeoffs of Different Assessment Strategies



February 2005

**Courtney L. Malloy, Ph.D. & Gwen C. Uman, Ph.D.**

In any educational or professional setting, making good decisions about competency is difficult, at best. Certification decisions – whether or not to certify an individual in a profession or trade – are particularly complex; organizations must create sound assessment procedures to ensure that appropriate decisions about the knowledge and skills of individuals are made. This brief discusses the advantages and disadvantages associated with the various types of tests used for assessing competency. We hope you find it useful as you and your organization develop and refine the testing procedures for your certification program.

## Assessment Strategies in Certification Settings

Three main types of assessment strategies are commonly used to make decisions about achievement and competency in certification settings: 1) Structured Response; 2) Constructed Response; and 3) Performance Assessments.

### Structured Response

This strategy is the most common type of strategy used by certifying organizations. In structured response assessments, the test taker is provided with a set of **pre-selected responses** from which to choose the correct answer. Tests are scored easily with the use of an answer key: either the test taker selects the correct answer or does not. The correct answers are counted to obtain the test taker's score. Many different types of test questions can be used in structured response assessments. Examples include:

- True/False questions
- Multiple choice questions
- Matching questions

### Constructed Response

**The test taker constructs** correct answers to questions in order to demonstrate mastery of content. Constructed response assessments require raters – or judges – to make decisions about whether the answer provided by the test taker is correct. Tests are generally scored using a rubric – a guide with the qualities, content, or processes the answer must contain in order to be correct. Examples of constructed response questions include:

- Essay questions
- Short answer questions
- Fill in the blank questions

### Performance Assessments

In performance assessments, the test taker responds to some sort of prompt (e.g., a written scenario, a live situation) that requires the **translation of knowledge into some sort of action**. In these assessments, the test taker must “perform” the skills that are required by his/her profession. Similar to constructed response assessments, performance assessments are often scored by raters using a rubric that details the attributes and procedures that must be present for successful demonstration of the skill. Examples of performance assessments include:

- Computer-Based Simulations
- Oral Questioning
- Live Skill Demonstrations

## Key Concepts: Validity and Reliability

All assessment strategies – structured response, constructed response, and performance assessments – have **both** advantages and disadvantages. Accurately measuring competency is tricky, and although we know a great deal about how to develop high quality measures, no test is perfect.

Before jumping into some of the tradeoffs associated with various assessment strategies, you should know about two key concepts that affect the quality of a test: **validity** and **reliability**. The higher the validity and reliability of a test, the better!

Organizations must create sound assessment procedures to ensure that appropriate decisions about the knowledge and abilities of individuals are made.

**Validity**

Validity is the degree to which a test measures the knowledge and skills it is supposed to measure. It is particularly important that the questions on a test adequately represent the various performance domains that are required to be competent.

For example, suppose you are interested in assessing whether someone will be a competent driver. A test with high content validity would likely assess things such as proper signaling procedures, making right and left turns, and braking at stoplights and stop signs. A test with low content validity might assess the tester's knowledge of car mechanics or the ability of the test taker to operate the radio and climate control functions in the car. Clearly the latter measures would not adequately assess the knowledge or skills required for driving a car effectively.

Now suppose you decide that a competent driver needs to know the following: laws governing driving, the meaning of various signs on the road, and how to operate a vehicle (e.g., braking, turning, accelerating). You develop a test for assessing competence. Seventy-five percent of the test you offer is related to the meaning of various signs on the road; 15% is related to how to operate a vehicle, and 10% is related to the laws governing driving. This test, although the content is related to the knowledge and skills required for driving likely has low content validity. The questions on the test are disproportionate to the actual knowledge and skills necessary for being a competent driver. The test probably does not adequately measure the extent to which the test taker can operate a vehicle or knows the laws and regulations necessary to be a good driver because only a small proportion of the test is dedicated to these two performance domains.

There are several ways to increase the validity of a test. Most relevant to certification programs is job analysis. A detailed job analysis provides the performance domains that are required for competence and the relative importance of those domains. For example, an analysis of the job "Personal Trainer," might indicate that the performance domains "Stretching Techniques," and "Business Management" represent knowledge and skills that are required for effective performance of the job. However, "Stretching Techniques" may be much more important (i.e., require more of a personal trainer's effort and time) than "Business Management." A job analysis would uncover that both domains are necessary, but also reveal that "Business Management" only represents 5% of a personal trainer's time whereas "Stretching Techniques" represents 15% of a personal trainer's time. A job analysis prior to test development increases the likelihood that the test questions accurately represent the performance domains **and** each domain's relative importance.

Each kind of assessment has various tradeoffs related to validity and reliability that certification organizations should know about.

**Reliability**

Reliability is the degree to which the results from one assessment would be similar if the assessment were administered again (with no additional education or training). In other words, a test is reliable when you would receive nearly the same score if you retake the test. If a certification candidate – let's call her Maria – receives a very different score the second time she takes a test, it is impossible to make a sound decision about her knowledge and skills. Which score should we trust – the first one or the second one?

A variety of factors can influence the extent to which a test is reliable. For example, there may be variations under which the test is administered. Maybe Maria was provided an hour to take the test the first time and only 30 minutes when she took the test the second time. Or, maybe the first test was administered in a quiet, controlled space and then in a noisy auditorium where construction was occurring – distracting Maria. If a constructed response strategy or performance assessment was used, reliability can be affected by two raters judging the same response differently. If two raters were used to score Maria's test, one rater may give her answers a higher score than the other rater. Reliability is also influenced by the number of test questions on an assessment. In order

for tests to be reliable, there must be a sufficient number of questions so that Maria has ample opportunity to demonstrate her knowledge and skills. When the number of test questions is limited, Maria's score can change dramatically depending on whether or not she knows the answers to the few questions asked.

Reliability can be increased by adding more questions to a test, by standardizing testing procedures, by developing rubrics that clearly specify the required responses, and by providing thorough training to raters.

**Advantages and Disadvantages of Assessment Strategies**

Each kind of assessment has various tradeoffs related to validity and reliability that you should know about before deciding what is right for you and your certification program. As you can see in the table on the next page, assessment strategies are not equally suited to every situation, and no one strategy is without potential problems.

**Structured Response** offers the opportunity to include many questions and test large numbers of test takers efficiently and cost-effectively. In addition, tests based on this strategy are easy to score: there is generally less opportunity for error in scoring than with constructed response or performance assessments. So, structured response assessments tend to be highly reliable. However, it can be time-consuming and expensive to develop valid tests that accurately measure all of the knowledge and skills required for competency in a profession. Structured response assessments often rely on simple recall and memorization. It can be difficult to test in-depth understanding and application of knowledge (although it is possible with well-constructed questions). Hence, if complex thinking or decision-making skills are necessary in the profession, structured response assessments may not always adequately measure competency.

**Constructed Response** provides the opportunity to easily test thinking skills and application of knowledge because test takers construct their own answers to questions. Constructed response assessments are powerful tools to use when you want to test in-depth understanding of a concept. In general, tests are also relatively less time-consuming to construct and revise. However, reliability can be difficult to attain with a constructed response assessment. Without clear criteria for scoring, raters may not judge responses in a similar manner. In addition, reliability may be affected if there are limited test questions. For example, essay tests commonly have only a few questions whereas multiple choice tests may have over a hundred. You also run the risk of assessing writing ability rather than the knowledge and skills you intend to assess.

**Performance Assessments** offer the opportunity to assess a test-taker’s potential reaction to a problem or situation experienced in the profession. They can also be used to test the hands-on decision-making skills that are required for effective practice. As a result, when designed well, performance assessments provide certification programs with excellent information about what a candidate would do in the “real world.” Similar to constructed response assessments, however, reliability can be difficult to attain with performance assessments. Poor rubrics and untrained raters can lead to inconsistencies in scoring. Moreover, there can be variations in test administration if scenarios are not standardized. For example, if actors play the role of “client” in scenarios that the test taker must respond to, variations in the actors’ behaviors or speech may influence test takers to respond differently – lessening the reliability of the assessment.

	Structured Response	Constructed Response	Performance Assessments
<b>Major advantages</b>	<ul style="list-style-type: none"> <li>Can administer several questions at each test administration (↑Reliability)</li> <li>Easy to score – not subjective, answers are correct or not correct (↑Reliability)</li> <li>Efficient and inexpensive to administer and score</li> </ul>	<ul style="list-style-type: none"> <li>Provides opportunity to test thinking skills and applications of knowledge (↑Validity)</li> <li>Easy to test in-depth understanding of a concept (↑Validity)</li> <li>Less time-consuming to construct and revise</li> </ul>	<ul style="list-style-type: none"> <li>Can create situations and/or settings that more closely resemble the real problems experienced in the profession (↑Validity)</li> <li>Provides opportunity to assess decision-making skills required for professional practice (↑Validity)</li> </ul>
<b>Potential problems to avoid</b>	<ul style="list-style-type: none"> <li>Poorly written questions – e.g., confusing questions, trick questions, more than one correct answer possible (↓Reliability, ↓Validity)</li> <li>Failure to sample performance domains representatively (↓Validity)</li> <li>Emphasis on recall and memorization rather than thinking skills or applications of knowledge (↓Validity)</li> <li>Time-consuming and expensive to develop</li> </ul>	<ul style="list-style-type: none"> <li>Poorly written questions – e.g., multiple interpretations of the same question by test takers (↓Reliability, ↓Validity)</li> <li>Possibility of measuring writing ability rather than knowledge (↓Validity)</li> <li>Poor scoring criteria – rubrics without clear guidelines (↓Reliability)</li> <li>Inconsistency among raters in scoring responses – raters might not judge the same response as correct (↓Reliability)</li> <li>Too few questions to adequately assess competency (↓Reliability)</li> <li>Time-consuming and expensive to score</li> </ul>	<ul style="list-style-type: none"> <li>Poor scoring criteria – rubrics without clear guidelines (↓Reliability)</li> <li>Inconsistency among raters in scoring responses – raters might not give the same score (↓Reliability)</li> <li>Too few exercises to adequately assess competency (↓Reliability)</li> <li>Variation in administration of prompts and test settings (↓Reliability)</li> <li>Time-consuming and expensive to administer and score</li> </ul>

## Avoiding the Pitfalls

The best way to overcome the problems associated with different types of assessment strategies is to use as many high quality tests as you can afford to use. **Multiple measures** of performance will yield the best and most dependable information about competency. Moreover, if you are making high stakes certification decisions – deciding that an individual is qualified to enter a profession, engage in certain practices, and work closely with clients – it is particularly important to have as much information as you possibly can about the knowledge and skills of candidates. More information will increase your ability to: 1) accurately assess the extent to which candidates can perform the required skills (validity!); and 2) make an appropriate decision about whether to certify or not certify (reliability!). Remember that when you include more test questions (or more measurements), reliability goes up! The likelihood that someone will pass if they passed the first time – or fail if they failed the first time – will increase as more measures are used. Also, using a variety of assessment strategies – for example, structured response and a performance assessment – will likely enhance your ability to adequately test all the skills needed for competency.

**Multiple measures of performance will yield the best and most dependable information about competency. If you are making high stakes certification decisions, it is particularly important to have as much information as you possibly can about the knowledge and skills of candidates.**

Keep in mind that one good valid and reliable test is better than two bad tests. So, whether you are able to use multiple measures or not, you should dedicate the time up front to develop fair, sound assessments. Here are a few tips to help you increase the validity and reliability of your assessments.

1. Invest time in a thorough job analysis. Although time-consuming, such work will deepen your understanding of the performance domains that should be tested as well as the relative importance of each domain.
2. Consult with subject-matter and measurement experts as well as certified practitioners. Subject-matter experts can verify that the content you are testing represents the required knowledge and skills. Measurement experts can help you develop and revise test questions and procedures to increase validity and reliability. Certified practitioners should serve on your certification board and be trained in test construction so they can write questions and judge incoming questions from other sources.
3. Develop expertise in test question construction. It is important to know how to write objective questions and answer choices. No trick questions! If you include trick questions, you will not be measuring the knowledge and skills you intend to measure.
4. Choose test questions that adequately represent all of the knowledge and skills you need to assess.
5. Standardize test administration practices. Eliminate sources of variation that may exist (e.g., testing locations, test instructions, materials, variations in performance assessment prompts, etc.).
6. Standardize scoring procedures. Develop clear, easy to understand rubrics. Train raters thoroughly on the use of rubrics to eliminate potential inconsistencies.
7. Always be ready to revise. New knowledge and skills often emerge that require certifying organizations to rethink test questions or the distribution of content on a test.

Making good decisions about competency can be difficult. However, if you use multiple measures and work to increase the reliability and validity of your assessments, you will make better certification decisions and strengthen your profession.

*Founded in 1982, Vital Research has expertise in research design, instrument development, psychometrics, qualitative and quantitative data analysis, and interpretation of results for immediate client use. We specialize in educational research, internal and external customer satisfaction research, service quality measurement, test development and accreditation services, and program evaluation.*



**6380 Wilshire Blvd., Suite 1609**

**Los Angeles, CA 90048**

**Phone: 888-848-2511**

**Fax: 323-653-0123**

**Visit us on the Web!**

**[www.vitalresearch.com](http://www.vitalresearch.com)**