

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

Interest in online testing is at an all time high. Fuelled by a revenue or savings opportunity and coupled with the availability of tools, trainers are designing, developing and delivering more online tests than ever before. This paper is recommended to anyone interested in competency-based tests of mastery, especially where an online, high-stakes, proctored design approach is required. Apply rigorous test design and development process and learn how online delivery will facilitate automated data collection and analysis methods. Make your online tests useful, credible, valid, reliable and legally defensible.

Consequences

Test developers use the terms low-stakes and high stakes to help describe a test's importance. Describing a test as high-stakes often creates a sense that the test has meaning, is difficult, requires more investment to develop, is electronically delivered and scored, or is developed by a trained psychometrician. While all of these characteristics are important they do not determine the test stakes. The only criteria that should determine the "stakes" of a test are the pass/fail consequences for the test sponsor, the test developer, and test candidate.

Test designers label their tests low, medium and high stakes using criteria that have little to do with test consequences and more to do with test design, development and delivery requirements or limitations. As an example, it would be wrong to label your test high-stakes because the target audience was globally dispersed. It would be wrong to label your test high stakes because you had a large budget.

Will the results of a test determine whether a candidate is hired or fired? Will passing an exam qualify an individual to perform tasks with a high degree of risk associated with them? What are the consequences of passing someone who is not qualified to perform the tasks covered by the test? Is there a need for workplace compliance monitoring? Could your organisation be sued as a result of a wrongful dismissal claim? Could the validity of your test be successfully challenged? Answering yes to some of these questions, is a good indicator that a test is "high stakes."

Using a test to make important decisions - raises the stakes and raising the stakes increases the requirement for development rigor. The reverse is not necessarily true. Even though a test may be considered low-stakes, it can still be a useful tool in determining a person's performance capabilities. A low-stakes test should be developed with the same rigor as a high-stakes test. No matter what decision you might make based on the test results, why would you ever create a test that was less useful, credible, valid, reliable or defensible? Identifying a test as low-stakes should simply mean the test results have will limited consequences

Imagine designing a test for the performance outcomes listed below and informally assign a low, medium or high-stakes label to each test:

- Remove a burst appendix.
- Describe emergency evacuation procedures to airline passengers.
- Provide directions from one end of Dubai to the other.

Deciding where your test falls on the continuum from low to high stakes is not a matter of preference. The criteria that should determine the "stakes" of a test are the pass/fail consequences for the test sponsor, the test developer, and test candidate.

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

- Design a reliable backup procedure for your office network.
- Clean and sanitise an elementary school washroom.
- Select an appropriate running shoe for a customer.
- Log and resolve a customer's product failure over the phone.
- Retrieve inventory from high shelves using a ladder.
- Perform exercises to reduce back and neck strain associated with desk work.

The tasks listed above all have varying consequences of succeeding or failing. Any test that accurately measured an individual's ability to perform the tasks would vary greatly in their design, development and delivery methods. Context and consequences help you evaluate the stakes, and knowing the stakes helps you determine the level of investment (effort, money, stakeholders) you need to accurately measure the associated skills and performance outcomes. Ultimately, pinpointing the stakes will help you refine your design, development and delivery strategy.

Context and consequences help you evaluate the stakes, and knowing the stakes helps you determine the level of investment (effort, money, stakeholders) you will need to accurately measure the associated skills and performance outcomes.

Proctoring

One of the important influences on test results is cheating. Will candidates cheat? Can they cheat? In most cases, given any stakes at all, candidates will cheat when they can. Certainly, the lower the stakes, the lower the motivation to cheat. When test security is low or the test is in circulation for a long time it is common to see test results rise as the test ages. People share test contents casually and the test is slowly "learned." So, even when the stakes are low, candidates still "cheat" inadvertently. The single and obvious response to cheating is proctoring and proctoring is the single biggest factor in driving test delivery costs upwards.

In most cases, candidates will cheat when they can. The single and obvious response to cheating is proctoring and proctoring is the single biggest factor in driving test delivery costs upwards.

It is hard to imagine someone being motivated to cheat on a test of how to remove a burst appendix. This maybe a case where the stakes are very high but motivation to cheat is low. In this situation, even the test candidate wants to ensure they are competent and would not risk performing the task without being qualified. In this case, there will likely be many test observers but these observers are acting as evaluators not as proctors. Many of the other listed tasks might however result in a temptation to cheat.

Would the retail shoe sales agent cheat on a product test? It might depend on the consequences of failing. Will the candidate be fired as a result of failing or would you simply recommend further product training if they failed. Would you cheat on a test of your knowledge of back and neck strain exercises? Clearly the stakes here are your own health so why cheat? Would you cheat on a test of your ability to resolve a specific customer problem? What if passing meant a new pay grade and better hours? Would you cheat on a test of your ability to navigate in Dubai? You might if your taxi driver's license was at stake. Just as it is easy to informally rate test stakes, it is easy to decide which tests would require proctoring. If the stakes are high enough, it is an unfortunate reality that cheating will occur and the test will warrant proctoring.

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

Scale

Audience size is the single important factor when deciding whether to deliver your test electronically. Unlike paper and pen tests, electronic tests will facilitate rapid test deployment, instant test scoring, automated data collection, improved data management and automated reporting. These benefits have varying effects on test stakeholders. Test sponsors can expect reduced delivery, data management and reporting costs. Test developers can expect automated test authoring tools along with rapid test delivery and scoring methods. Test candidates can expect instant test scoring and reporting.

Electronic test delivery is not, however, a panacea for your assessment challenges. There are drawbacks and limitations to electronic testing. Although not impossible, developing multiple-choice items that capture higher-level cognitive thinking is more difficult and time consuming. Electronic simulations, role plays and graphical items which can capture high-order thinking are expensive to develop and score. Weigh the pros and cons of electronic testing (higher investment but easy data management) and weigh the pros and cons of paper and pen tests (low investment but manual data collection, scoring and management).

If the audience and budget are small, you would be well advised to stick with paper and pen and invest your money in the test development rigor rather than an electronic delivery method. Spending money on a slick way to deliver your test is not a solution and speed of delivery, scoring and data management is only an advantage when you haven't sacrificed test design and development effort. Like electronic testing, paper and pen tests presents their own set of pros and cons.

In many cases, test developers will take advantage of the human scoring required by paper and pen tests and make use of open-ended questions (essay response). The advantage of open-ended items is it is easier to develop an item that targets higher-order cognitive competencies. Scoring open ended questions is usually slower because it requires human interpretation but human interpretation means more evaluative flexibility. Test developers are less likely to include open-ended items in their electronic test since a primary objective with electronic testing is the elimination of human scoring.

Collecting hard copy tests and scoring them is very work intensive and time consuming, but their use is based on the assumption that the audience size is small so your effort should be manageable. Scoring templates will speed up the process and take some of the work out of scoring multiple-choice tests. Human scoring also presents challenges around consistency of scoring. This is especially true of live observation where scoring schedules are a requirement to help create consistency across observers. Variability in scores is a validity and reliability killer and you don't want to introduce variability through your scoring method.

Proctoring and electronic delivery will create incremental testing costs but these costs are warranted if the audience is large and the stakes are high. Delivering a proctored test globally does present logistical challenges but there are solutions available with major vendors such as Sylvain Prometric and Thompson VUE ready to deliver tests to a dispersed audience through their global network of test centres.

Electronic test delivery is not a panacea for your assessment challenges. You must weigh the pros and cons of electronic testing. Decide between the higher investment and easy data management of electronic testing or the low investment and slower data collection and management of paper and pen tests.

Proctoring and electronic delivery will create incremental testing costs but these costs are warranted if the audience is large and the stakes are high.

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

If the audience is geographically dispersed but not necessarily large you will be faced with an alternate logistical challenge and potential higher per-test delivery costs. In these cases, you may need to seek creative solutions or compromises. As an example, if you must proctor a paper and pen test for a small audience, you might consider administering the test immediately following classroom training since there is an instructor in the room and the audience is assembled.

The consequences (stakes) of a test will determine the need for proctoring. The audience size, which is often linked to budget, will determine the benefits of electronic testing. The quick advice is to proctor tests only when absolutely necessary and use electronic delivery methods if the budget warrants their use. Deciding to use proctoring and electronic testing are the first two elements in determining the design of your test. With proctoring and delivery mode decided, other design, development, delivery, and data management elements will fall into place.

The consequences (stakes) of a test will determine the need for proctoring and the audience size, which is often linked to budget, will determine the benefits of electronic testing.

Rigor

Once you have decided on the need for proctoring and decided on a test delivery method, you can go about designing the test, writing items (questions), delivering the test to the target audience, scoring the test, and reporting the results. In a nutshell, this is the development process but the stakes and delivery method will play a role in determining the exact steps you should follow. All tests must be useful, credible, reliable, valid and defensible and to achieve this end, you must exercise a certain amount of rigor. Assuming you are committed to a rigorous process and electronic delivery is the right solution for you, the remainder of this whitepaper will provide guidance on how to design and develop an electronically delivered, defensible test with multiple-choice and multiple response item types.

Generally, it is bad practice to assign this task of developing a high-stakes, competency-based test of mastery to subject matter experts. While they may know a lot about the test content domain, they are not necessarily qualified to plan, manage and deliver the right results given the consequences of passing and failing the wrong people. Not all experts make good teachers and not all teachers necessarily make good test authors. Experts often under and over estimate the capabilities of a target test audience and do not always understand the broader context of a test.

Not all experts make good teachers or test authors.

The primary goal of any test is to distinguish between those that “know” and those that “do not know.” This means the test should be fair, questions should be easy to read and people that “know” should be able to pick the right answers quickly. People that don’t know should find it equally difficult to choose between any of the question options. The test should never trick the test taker since you may end up tricking the wrong people. Your biggest concern will be that the wrong people pass (false positive) and fail (false negative) the test. Although most people worry more about false positives more than false negatives, be concerned with both. The credibility of your test will be challenged if people that “know” - keep failing.

The primary goal of any test is to distinguish between those that “know” and those that “do not know.”

Create a Test Definition Document. This document will serve as the foundation for the test development, providing details on the test purpose, content domain, target audience, delivery method, consequences, and outcome objec-

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

tives. The Test Definition should include a description of the method used to establish a cut score (pass/fail) for each test. Common methods include Anghoff and the Borderline Survey Method. When multiple test of the same type are planned, create templates for the Test Definition Document to ensure consistency and lessen the work load.

Once you have a complete Test Definition Document and before developing any test items, you should develop a test Blueprint. Using feedback and advice from subject matter experts, create a comprehensive list of all test objectives and then organize the objectives into test sections. Report the breakdown of question percentages (weighting) and quantities in the Blueprint and then use the Blueprint as a item development planning tool.

Develop your test items using the Blueprint as your guide. As you create test items, assign them to their associated test objectives within each objective and section within the Blueprint. Keep all items in one place, creating a central repository of the items called the Item Bank. The Item Bank might consist of a Microsoft Word Document or a Relational Database. The format of the bank will likely be determined by your test publishing method. If publishing the test through the Sylvain Prometric test network or through Pearson VUE® Authorized Center (PVTTC), you will be asked to submit your bank for publication in an accepted test format.

Now that you have an Item Bank, assemble a group of subject matter experts to review each item for technical and cognitive congruence with their associated test objective. The purpose of this Technical Review Workshop is determine each item's eligibility for inclusion in the Beta Item Bank. Again, to ensure consistency and lessen the workload, you should develop a set of item development standards that will be used when authoring and reviewing all exam items.

Item writing rules must be enforced consistently and all items must be diligently reviewed for technical accuracy and consistency. The item writing rules will help ensure that items are not tricky, that they are at the right cognitive level, that they use clear language and that each item is concise and accurate. The most basic set of rules would include:

- Ask a question
- Place the details of the question in the stem
- Make the options as short as possible
- Make all distractors (wrong answers) plausible
- Make the right answer(s) clearly right and the wrong answer(s) clearly wrong.

Remember the goal is to divide between those that know and those that don't know. You want to make it easy for the competent people to pass your test and you don't want incompetent people to be able to rely on guessing or test wiseness.

Before developing any test items, you should develop a test Blueprint.

Have subject matter experts review all items to ensure the technical and cognitive congruence and accuracy of the items.

Data Collection & Analysis

Measure Twice, Cut Once

Designing and developing competency-based tests of mastery for online delivery.

Whether the stakes are high or low, you should perform basic statistical analysis on your items before publishing the test to your target audience. Any claims you make about the reliability and validity of your test should be backed by evidence and should be established using accepted psychometric methods. If the test is challenged, the rigor of the process, the involvement of key stakeholders in the process, and the empirical evidence will all combine to provide your defensible position.

Administer the entire Item Bank Item to a sample audience. The Beta Test will provide important data regarding each item's difficulty and ability to discriminate between those that know and those that don't know. Further data analysis will establish overall ratings of the test reliability and provide evidence for a calculated pass/fail cut score. These data analysis tasks are best assigned to a trained psychometrician.

Trained psychometricians will help you establish comprehensive statistical analysis methods that will ensure that all exams are valid and reliable. Performing this level of analysis will guarantee that your test is defensible but more importantly that the test is a true measure of the prescribed performance competencies. Here are key data analysis tasks:

Item Analysis	Calculate Item Difficulty and Item Discrimination statistics to be used in Item Selection and then make final item selections using the data.
Form Balancing	Form balancing ensure that each version of the test can be considered equal and will produce the same test result.
Reliability	Estimate test reliability (test-retest & internal consistency) using accepted measures such as the Juder-Richarson formula, split-half method or Cronbach's Alpha
Cut-Score	Establish the test cut-score using an accepted method such as Anghoff or the Borderline Group method. The borderline group cut-score method compares the median test score of candidates in a minimally-qualified group with their actual test scores to establish a suggested cut score.
SEM	Calculate the Standard Error of Measurement (SEM) to determine the amount of test score variation. The SEM will be a factor in determining the cut score for the test.
Reporting	Collect data and generate detailed reports of the Item Analysis, Item Selection, Cut Score Analysis, Form balancing, and Test Reliability.

Plan, design, develop, analyse and deliver your competency-based test of mastery. From Test Definition Document to a published test, this process should take between eight and 12 weeks for a test of 50 multiple-choice questions. This schedule will depend on the length of the test (number of items per form), the number of forms (versions of the test), resource availability (especially SMEs) and the length of your Beat Test period (typically three to four weeks).

Informal or anecdotal affirmation that the test is valid and reliable is not a defensible position. If you want to make sure your test is consistently measuring what you want it to measure, perform basic data analysis to ensure the test meets accepted standards.

About the Author

Roy Lamond began his career in adult education in 1988 as an ESL teacher, Roy moved to technical training, working throughout North America as an independent Microsoft Certified Trainer during the mid-90s. More recently, Roy has held positions as a Director of a Technical College and President/Founder of an eLearning development firm in Canada. Holding a Master of Education from the University of Ottawa in Educational Measurement, Roy's primary interests are computerized assessment and performance evaluation. Roy has worked with global organisations such as Symantec Corporation, Business Objects, Genesys Labs, OZ Communications and Research in Motion to develop high-stakes technical certification programs.

