Assessments through the Learning Process

This paper defines different types of assessments and explains their respective applications. It explores how instructors and organizations can use assessments to improve the learning process. It is designed to help readers distinguish among different types and styles of assessments and understand various assessment tools. Readers will also learn how to develop effective assessments and analyze their results.

Authors: Eric Shepherd
Janet Godwin

With assistance from: Dr. Will Thalheimer, Work-Learning Research
Dr. William Coscarelli, Southern Illinois University
Dr. Sharon Shrock, Southern Illinois University
Brian McNamara
Joan Phaup

Table of Contents

Introduction

Assessments – quizzes, surveys, tests and exams – are everywhere.

In addition to helping learners, assessments are used extensively to qualify people for various positions or job roles, which they subsequently perform. Throughout the processes of learning, qualifying and performing, organizations accumulate evidence and feedback that help in decision making. Much of this data is collected through the use of assessments, which are widely used for talent acquisition, job task analysis, talent management, performance measurement, and many other organizational processes. This paper, however, will focus on assessments related specifically to the learning process.

Learning, education and training are now delivered in many ways including classroom instruction, televised lectures, online courses, podcasts, and short on-the-job learning modules. Whatever the delivery method, it's critical to understand how people learn, what they have in fact learned, and whether this knowledge is useful for their particular role. The cornerstone of developing successful educational, training, and certification materials is the effective use of assessments. Learning experiences are enhanced when they include the search and retrieval practice afforded by quizzes, tests and other assessments. This retrieval practice is an essential factor in determining what learners actually remember and apply.

Well-crafted assessments can guide people to powerful learning experiences; reduce learning curves; extend the forgetting curve; confirm skills, knowledge and attitudes; and motivate people by giving them a solid sense of achievement.

The purpose of this white paper is to illustrate how both organizations and instructors can use assessments to improve the learning process and achieve greater results. The paper is designed to help readers:

- Distinguish among the types and styles of assessments
- Distinguish among the various types of assessment tools
- Know how to develop effective assessments
- Know how to analyze the results of assessments
- Understand reliability and validity
- Understand the benefits of computerizing assessments

1. An Introduction to Assessments

It is important to define the context of assessments in the learning process. There are many styles of assessments that are not dealt with in this paper, such as medical assessments by a doctor, risk assessments in hospitals and assessments for the accreditation of colleges and universities, to name a few. In this paper we use the generic term assessments to describe quizzes, tests, surveys, and exams. These instruments assess learners' knowledge, skills, abilities and attitudes.

The table below further defines these terms as used in this white paper:

Assessment	Any systematic method of obtaining evidence by posing questions to draw inferences about the knowledge, skills, attitudes, and other characteristics of people for a specific purpose.
Exam	A summative assessment used to measure a learner's knowledge or skills for the purpose of documenting their current level of knowledge or skill.
Test	A diagnostic assessment to measure a learner's knowledge or skills for the purpose of informing the learners or their instructor of their current level of knowledge or skill.
Quiz	A formative assessment used to measure a learner's knowledge or skills for the purpose of providing feedback to inform the learner of his or her current level of knowledge or skill.
Survey	A diagnostic or reaction assessment designed to measure the knowledge, skills, and/or attitudes of a group for the purpose of determining needs required to fulfill a defined purpose.

1.1 The Uses of Assessment

There are five primary purposes or uses of assessments as described in the table below:

Diagnostic	An assessment that is primarily used to identify the needs and prior knowledge of participants for the purpose of directing them to the most appropriate learning experience.
Formative	An assessment that has a primary objective of providing practice for search and retrieval from memory for a learner and to provide prescriptive feedback (item, topic and/or assessment level).
Needs	An assessment used to determine the knowledge, skills, abilities, and attitudes of a group to assist with gap analysis and courseware development. Gap analysis determines the variance between what a learner knows and what they are required to know.
Reaction	An assessment used to determine the satisfaction level with a learning or assessment experience. These assessments are often known as Level 1 evaluations (as per Dr. Donald Kirkpatrick), course evaluations, or smile or happy sheets; they are completed at the end of a learning or certification experience.
Summative	An assessment, usually quantitative, with the primary purpose of giving a definitive grade and/or make a judgment about the participant's achievement. If this judgment verifies that the participant has met an established standard indicative of special expertise, the judgment may confer "certification."

1.1.1 Diagnostic Assessments

If you go to the doctor and just say, “I have a pain,” you’d be concerned if the doctor says, “Oh, here’s a pill.” You’d feel far more confident if the doctor asks, “Where is the pain? How often does the pain occur? Have you done anything recently that might have caused this pain?” The answers to these questions help the doctor tease out the issues so that he or she can make a diagnosis and write a prescription. That’s exactly what happens during diagnostic assessments.

Diagnostic assessments are typically used in pre-learning assessments, before a person engages in a learning experience or a placement test. For example, a college student whose second language is English might take a test to discover if his or her English skills are adequate for taking certain courses. The test measures that person’s current knowledge and skill, thereby helping the instructor tailor the course effectively. These kinds of test also create intrigue, which can in turn actually enhance the learning experience. For instance, if an instructor asks a question that a student can’t answer, that student might become curious to find out the answer and therefore pay more attention in the class.

Diagnostic assessments are used to determine knowledge and identify skills gaps and needs. Such an assessment might report that a learner has mastered every competency in using Microsoft Word but can only perform 50 percent of those required to use Excel. The results of the assessment would prescribe a course on Excel. In addition, this type of assessment can place students within suitable learning experiences by asking diagnostic questions such as, “Do you prefer instructor-led training or online training?”

1.1.2 Formative Assessments

Formative assessments provide feedback to individuals and their counselors during the learning process by providing search and retrieval practice. When people must provide answers to questions about material they’ve learned, their brains must search their memories and retrieve the information. These memory processes help solidify the learners' knowledge and help maintain that information in an accessible state for later recall. If a person answers incorrectly, the instructor now has a teachable moment or an opportunity to provide feedback, and says, “No, that’s not quite right…this is really the right answer” or “No, but have you thought about the problem this other way…”

Search and retrieval practice is often used for:

- Practice tests and exams
- Self-assessment of knowledge, skills, and attitudes for the purposes of learning

Formative assessments help reassure learners that they’re actually learning or alert them that they are not learning and provide feedback to correct any misconceptions. Research on Web sites has found that people tend to take quizzes first and use the feedback so they can say, “Hey, I’m doing pretty well in this subject. I’m going to move on,” or “I need to study this topic more.” Not only did they discover their level of competence, but they also inadvertently reduced their forgetting curve by experiencing some search and retrieval practice. These formative assessments are sometimes used to collect data that contribute to overall grades. They’re not like the final exam, but are rather like a series of small tests that provide evidence that helps instructor make a judgment.

1.1.3 Needs Assessments

Needs assessments evaluate the knowledge, skills, abilities, and attitudes of a group to help someone determine the group’s training or provide data for a job task analysis. These are low stakes assessments that measure against requirements to identify a gap that needs to be filled. These assessments allow training managers, instructional designers, and instructors to work out what courses to develop or administer to satisfy the needs of their constituents.

1.1.4 Reaction Assessments

A reaction assessment occurs when we obtain learners’ reactions and opinions about their learning experience. This is typically referred to as a smile sheet, and under the model developed by Donald Kirkpatrick, it’s referred to as a level 1 evaluation. In colleges and universities, it’s typically referred to as a course evaluation. Such an assessment gathers opinions from the students about what they thought of the course materials, the instructor, the learning environment, the directions to the facility, audio-visual support and so forth. From the information gathered, the instructor can improve the learning experiences going forward.

1.1.5 Summative Assessments

Summative assessments are just what they sound like: they sum up the knowledge or the skills of the person taking the test. This type of assessment provides a quantitative grade and makes a judgment about a person's knowledge, skills and achievement. Summative assessments include post-course exams, licensing exams, certification tests, and other medium or high-stakes assessments—both regulatory and non-regulatory. These assessments provide quantitative scores signifying how much knowledge or skill the participants have acquired.

1.2 The Stakes of an Assessment

Before examining how assessments can most effectively be used in the learning process, it's important to understand that the various types of assessments can be categorized in terms of high, medium or low stakes.

The level of an assessment's stakes refers to the consequences to the candidate. For example, an exam normally has a higher consequence, while a survey has low or even no consequence.

In low-stakes assessments such as quizzes and surveys, the consequences to the candidate are low, and so the legal liabilities are low. These assessments require less planning than high-stakes assessments: Subject matter experts (SMEs) simply write the questions and make them available to learners. Low-stakes assessments are often taken alone since there isn't any motivation to cheat or share answers with others. Therefore, proctoring or invigilating is not required. This means that test administrators would not normally check the ID or watch someone taking a low-stakes assessment, whereas they would in the case of a high-stakes exam.

High-stakes tests require a lot of planning because they must be valid and reliable. They may require job task analysis, setting the pass/fail scores, specifying the methods and consistency of delivery required, and determining how results will be stored and distributed. Job task analysis discovers what tasks are associated with the job, how often they are completed and how important they are to the job. Test developers must carefully plan which questions should be in the test by topic, which subjects are of more and less importance, and the depth of competency required. The pass/fail score or cut score determines the threshold between passing and failing. Security measures must be well thought through as well, in order to protect the content of the test and to protect against cheating.

The general rule for a test or exam is that in a work setting it should look like the job; at an academic institution the test or exam should look like the curriculum.

Finally, in a high-stakes assessment, psychometricians will analyze the resulting statistics and provide guidance on how to improve the wording of questions, the wording of choices, or how to improve the overall test. In a low-stakes assessment, however, it's rare to involve psychometricians.

1.3 Factors Determining the Stakes of an Assessment

The stakes of an assessment also determine a number of other factors, from the overall consequences to the validity of the test itself.

	Low	Medium	High
Consequences	Few	Some	Major
Decisions	Few and easily reversed	Can be reversed	Very difficult to reverse
Options for participant	Refine studies	Pass, fail, or work harder	Pass or fail
Motivation to cheat?	Low	Medium	High
ID individual	Not important	Potentially important	Very important
Proctoring required	No	Sometimes	Always and constant
Development effort	Minor	Medium	Major
Check reliability and validity	Rarely	SME	Psychometrician

The chart above shows how the consequences of different types of assessments vary. A high-stakes exam might determine whether or not a person is hired or fired or will graduate from college. As would be expected, decisions in response to low-stakes tests are few and easily reversed. If someone gets a poor score on a quiz, they can very easily appeal it, but if they get a failing score on a nursing certification exam, appealing it would be very difficult, if not impossible. The options for the participant vary in direct relation to the stakes.

There is relatively no motivation to cheat on a survey and very little to cheat on a low-stakes quiz. Quizzes are really learning aids, so a person would only be cheating themselves. However, on a nursing, architectural or engineering exam, the stakes are much higher. There is a strong motivation to cheat, so it is important to identify each test taker. For high-stakes tests related to national security, government agencies or the military might use biometric screening such as retinal scans to confirm test takers' identities.

If there is a low motivation to cheat, there is no need to proctor an assessment, but if there is a high propensity or motivation to cheat then there should be attentive, continuous proctoring.

The development effort for a quiz is quite minor. But for medium and high-stakes assessments, a subject matter expert is likely to develop an average of three questions an hour, and of those, only one will be included in the test. High-stakes questions take far more time to develop. The average cost for an SME to develop a question that gets into a high-stakes exam, is between $500 and $1,500! It's important to make sure that each question resonates with the whole test, that it's valid, and that good, more knowledgeable candidates tend to get it right while poor, less knowledgeable candidates tend to get it wrong. Consequently, it takes time, effort and thought to get the right mix of questions into a high-stakes test or exam.

1.4 Assessment Applications

Each type of assessment may be mapped to typical uses and stakes as outlined in the table below:

Assessment Type	Assessment Use	Assessment Stakes
Exam	Summative	Medium, High
Test	Diagnostic	Low, Medium
Quiz	Formative	Low
Survey	Needs, Reaction, Diagnostic	Low

In turn, each type of assessment has numerous applications in real life:

Surveys	**Main purpose of assessment is to gather information or opinion**
Level 1 Surveys, Course evaluations	These are specific evaluations, given after a course, to get the student's feedback on the learning activity. Level 1 Surveys are also called course evaluations or smile sheets. Level 1 Surveys typically ask for the learner's evaluation of the course content, course materials, the instructor, the information delivery, etc.
Needs Analysis Surveys	These surveys are given to a group to explicitly determine what areas they want to learn more about, or where their knowledge is weak, with the purpose of providing learning offerings to meet those needs.
Job Task Analysis Surveys	These surveys are given to a group of people doing a job to determine which tasks they perform, how regularly they perform them, and the significance of the task to their role.
Employee Attitude/ Opinion Surveys	These surveys are used within an organization to collect input from employees about their feelings on one or more topics.
Customer/Partner Satisfaction Surveys	These surveys are used by an organization with other groups than their employees; it might include customers, prospects, partners, and others outside the organization. Classic market research is an example of this category.
Information and Opinion Surveys	These surveys are used to gather information and opinions that do not fall into the categories above. For example, surveying conformance with maturity models or six sigma.
360 (Level 3) Surveys	These are surveys designed to measure whether behavior has changed on the job, usually delivered by a 360/180 survey to colleagues. Often data is aggregated to identify the effectiveness of learning activities.
360 degree survey (employee appraisals)	Information gathered from someone's self-assessment and assessment by peers (and where relevant, superiors and subordinates) to obtain input on how the person can improve. Organized by HR as part of a performance appraisal.
360 learner peer review assessments	Information or judgments gathered from peers in a learning context to aid learning and provide feedback.
Formative assessments	Strengthen memory recall through retrieval practice, correct misconceptions and promote confidence in one’s knowledge
Quizzes during Learning	Quizzes are typically used at the end or at intermediate points of learning activities. Usually the results are not stored (or if they are, it is just to see if questions should be promoted or for other very low-stakes use).
Practice tests	Practice versions of higher-stakes exams, made available or sold for practice purposes.
Diagnostic assessments	Assess knowledge, skills, behaviors, understandings, and/or attitudes to determine gaps and potentially provide a diagnosis and prescription of learning and/or other activities
Pre-Tests	A diagnostic assessment before a specific learning activity. Used to create intrigue, to set a benchmark for comparison with a post course test, as a pre-requisite or route to an appropriate learning activity, and to provide instructors and mentors with information on the student's abilities.

Placement tests	A diagnostic assessment used as a measure to place someone in the right course or learning activity. Has a summative element, but its main purpose is to diagnose and prescribe. Used as a route to the correct learning activity.
Self-diagnostic tools	Assessment that asks non-judgmental questions and gives feedback to the participant to provide recommendations for products, services and/or learning activities. An example might be a financial survey that asks for information and recommends financial products.
Personality assessments	Assessment that analyzes personality traits in order to predict behaviors, including the Myers-Briggs. Sometimes gives information directly to participant and sometimes requires an expert to review and interpret.
Summative tests	Primarily measure or certify knowledge, skills and aptitudes (KSAs)
Post Course Tests	Post Course Tests (aka Post Tests) are typically given to identify the learner's KSAs and may contribute to passing the course. Sometimes data from these tests are used to indicate how much a student or group of students have advanced their knowledge, skills and attitudes through a comparison to the pre-test.
Exams during study	Assessments given during or at the end of a prolonged course of study that are used to define the grade or passing of an extended learning activity, for instance university courses. Exams are normally available only to students who have registered with an organization and have been studying with that organization, often in a cohort. Are most commonly academic, e.g., university exams.
Internal exams	Exams open to a closed community, for instance employees or partners set by an organization for its own purposes. Examples include regulatory compliance exams and partner verification exams.
Open exams	Exams open to anyone who qualifies, used to certify or score KSAs. Common examples are IT certifications (e.g. Microsoft, Cisco) or general academic entry exams (e.g. TOEFL, SAT) and IQ tests. Authority comes from an organization, not a legal source, but passing can be a passport to greater pay or rewards and so the stakes are often high.
Licensing exams	Exams that must be passed to meet a legal requirement or a quasi-legal requirement, for example driving tests, medical licensing exams. These are very high stakes and need to be legally defensible.
Pre-employment screening	Asks questions of applicants about KSAs and experience to gate applicants. Used to determine whether applicants meet basic requirements and can be considered for the next stage of a recruitment process. Are in some cases close to a survey.
Psychological assessments	Clinical psychological assessments that ask questions and then determine a psychological profile. These are typically validated against a norm population.
Pre-employment tests	Tests given in pre-employment to measure KSAs to see if the applicant is suitable for recruitment.

1.5 Consumers Have a Stake in Assessments

When a plumber comes to your house, will he break a pipe? Is a surgeon qualified to take out someone's gall bladder? Can you trust your laptop to a service center technician? Should your teenager have a driver's license? Will an airline pilot respond skillfully to an emergency? While we might feel sorry for someone who fails a test, our greater concern is whether we can trust someone to perform a particular task or job competently. So a kind of partnership has developed among consumers, test takers and test designers. Consumers want to know that they can trust the people they hire;

designers of high-stakes tests want to measure accurately with reliable and valid tests; and candidates want the tests to be fair.

There must be communication at each level to ensure that everybody understands that designers are trying to produce a fair, workable system for an assessment. These issues are related to Face Validity, which is the perception among non-experts that a test measures what it is intended to measure. Not only does a test need to have Content Validity, the documented determination by experts that the test fairly measures the desired job competencies, but it also needs to be trusted by the consumer and the candidate taking the assessment. The test developer needs to be trained properly; the participants need to be educated about the value of an assessment and be reassured that they're ready for the experience; and the consumers need to be educated about the validity of the assessment so they can trust the people doing their work.

2. The Reliability and Validity of Assessments

An assessment is reliable when it measures the same thing consistently. If a survey indicates that employees are satisfied with a course of instruction, it should show the same result if administered three days later. (This type of reliability is called test-retest reliability.) If a course instructor rates employees taking a performance test, their scores should be the same as if any other course instructor scored their performances. (This is called inter-rater reliability.) An assessment is valid if it measures what it is supposed to measure. If a test or survey is administered to happy people, the results should show that they're all happy. Similarly, if a group of people who are all knowledgeable are tested, the test results should reveal that they're all knowledgeable. Good assessments are both valid and reliable.

If a job-related assessment is valid, it looks like the job, and the content aligns with the tasks of the job in the eyes of job experts. This type of validity is known as Content Validity. In order to insure this validity, the persons creating the assessment must first undertake a job task analysis to analyze what tasks are required for a particular job. They do this by surveying subject matter experts (SMEs) or people on the job to determine what knowledge and skills are needed to perform job-related tasks.

Test validity requires test reliability. However, a test can be reliable but not valid. An example illustrates how the reliability and the validity of an assessment are related. If we gave a vocabulary test twice to a group of nurses, and the scores came back exactly the same way both times, the test would be considered highly reliable. However, just because the test scores are reliable does not mean that the test measures nursing competence. The test is reliable, but it is invalid as a measure of skilled nursing. It is merely a consistent measure of vocabulary. Now imagine that a test of nursing skills is administered twice to a group of skilled and unskilled nurses and the scores for each examinee are different each time. The test is clearly unreliable. And if it's not reliable, it cannot be valid; fluctuating scores for the same test takers cannot be measuring anything in particular. So the test is both unreliable and invalid. The reliable and valid test of nursing skills is one that yields similar scores every time it is given to the same group of test takers and discriminates every time between good and incompetent nurses. It is both consistent and measures what it is supposed to measure.

Another more visual example of the relationship between reliability and validity is represented by the dart boards in Figures 1, 2 and 3.

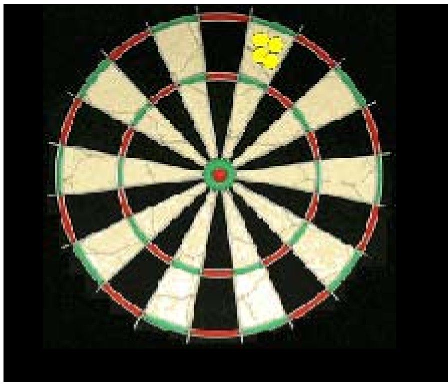


Figure 1 Reliable (Consistent) but not Valid

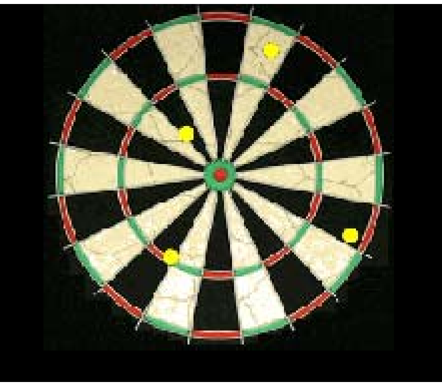


Figure 2 Not Reliable (Consistent) and therefore it cannot be Valid

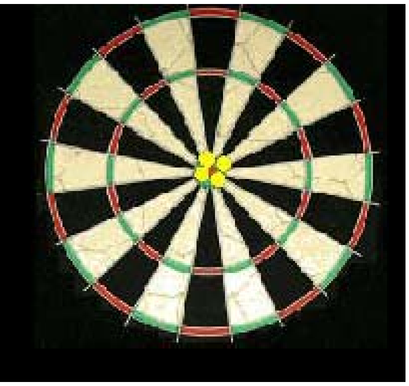


Figure 3 Reliable and Valid

The dart board in Figure 1 shows that all the darts are stuck in the same area, illustrating that the thrower – the analogue of an assessment score – is reliable and consistent, but unfortunately his throws are not valid. If his throws were valid, all the darts would be in the center, the bull's-eye. In Figure 2, the darts have landed all over the board. This assessment is not reliable because it is not consistent. Finally, the last example is of an assessment that is both reliable and valid, because all of the scores are clustered together and on target with the job. Notice that reliability is possible without validity, but that validity is impossible without reliability.

2.1 Assessment Score Interpretations

When people take an assessment, it's important for them to understand the implications of their scores, particularly when passing or failing make a major difference in their lives. There are, generally speaking, two ways to score an assessment. These are referred to as 'criterion-referenced' and 'norm-referenced'. Pass scores for norm-referenced tests are determined after a test is completed, while pass/fail scores for criterion-referenced tests are determined beforehand.

With a criterion referenced score interpretation, the test designers use a standard-setting process to determine an acceptable standard for setting the pass or fail score. If someone passes this test, they are determined to be qualified, whether it's as a surgeon or a plumber – whatever job competencies the test measures. Figures 4 and 5 illustrate mastery curves for these two different types of assessments.

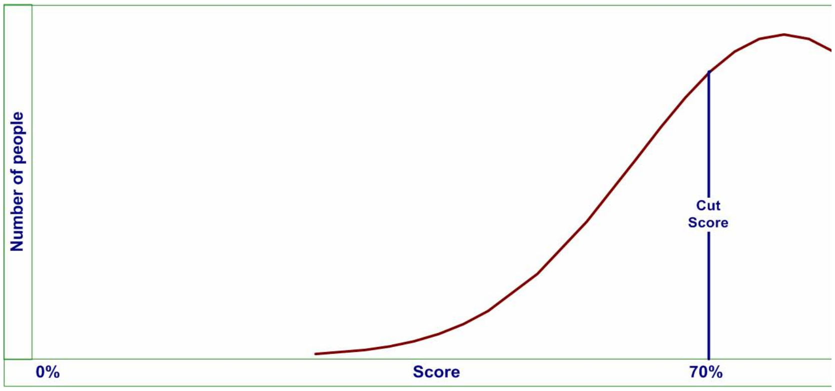


Figure 4 Typical mastery curve for a criterion referenced test

This curve shows the number of people who took the assessment and the scores they achieved. The bottom scale goes from test scores of zero up to 100, while the left hand side of the scale shows the number of people who achieved a particular score. The cut score has been determined to be around 70 percent, which was probably set by subject matter experts who had determined the competencies required to pass the exam.

With a criterion-referenced score interpretation, the number of people who will qualify from examination event to examination event is irrelevant, since each sitting will include candidates with more or less knowledge. What's important, however, is that a benchmark has been established for the standards required to adequately perform a particular job. As an example, a driving test will use a criterion-referenced score interpretation, as a certain level of knowledge and skill has been determined to be acceptable for passing a driving test.

A norm-referenced test, on the other hand, compares the scores of an examinee against the scores of other examinees. Often, typical scores achieved by identified groups in the population of test takers, i.e., the so-called norms, are published for these tests. Norm-referenced tests are used to make "selection decisions." For example, a college entrance exam might be designed to select applicants for 100 available spaces in a college. College decision makers might use the test scores to determine the 100 best people of those taking the test to fill those positions. In some years a higher quality group of students will qualify and sometimes a lower quality group. The key, however, is that the test will spread the examinees' scores out from one another so that the 100 best performers will be readily identifiable.

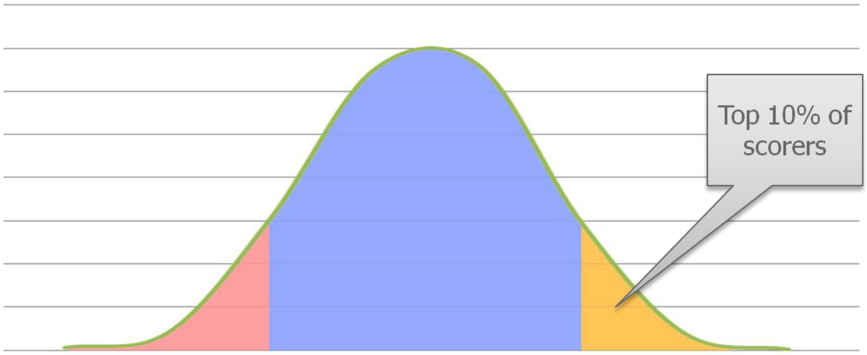


Figure 5 A norm-reference test is designed to "spread" and rank participants within a group so to easily identify top performers

How are these test philosophies important? If a city decided to commission an architect to design a building, the planning commission would want to make sure that the architect had passed a test that was criterion-referenced. They wouldn't want to commission someone to undertake a large engineering project based on the fact that they were one of the best from the class of '77. On the other hand, a norm-referenced test could select the top 10 sales representatives or honor students of the year.

As consumers we feel comfortable knowing that our doctors, nurses, and pharmacists have passed a certification exam that determined that they were competent and had the required knowledge and skills for their job. It would be disconcerting to learn that your doctor had graduated from an unknown university that always certified the top 50 students regardless of their abilities.

2.2 Timed, Speeded and Power Assessments

Most tests are timed in some way, but in general the time limits are set so that 95% or more of students finish the tests in the designated time frame. Speeded tests – which might be used to test clerical personnel using fairly basic content — are used to primarily to measure speed of performance. The score might be expressed as the number of correct answers provided within a given time limit. In contrast to speeded tests, power tests measure capability to perform correctly regardless of speed of performance, although for practical reasons, all tests have some imposed time limit. It is important to note, however, that in criterion-referenced testing the goal is to measure a specific real-world competency, and in the real world, both power and speed often are essential to satisfactory performance.

For example, an item within a criterion-referenced test for a nuclear reactor room technician might simulate a dangerous situation by sounding alarms and/or displaying graphics of control panels. This stimulus requires the person to act within a certain timeframe. The situation calls for immediate action and doesn't allow the person to consult with job aids to determine the best course of action. The person must know what action to take and do so within a given time limit. The content of a criterion-referenced test is driven by the conditions and behaviors contained in the objectives or competency descriptions that are the foundation of the test. Those objectives often combine both the need for correct action and the need for speed in executing those actions.

3. Designing Appropriate Learning Experiences

Assessments provide a valuable tool for organizations to properly design learning experiences so they are effective and useful. Doing so involves a five-step process:

1	Define objectives	Does a company want to increase customer satisfaction, reduce error rates or improve the safety record of a factory? What are the learning objectives of a given college course? What do students need to do to demonstrate that they have mastered a particular subject? Defining objectives is key to developing a relevant, effective learning experience.
2	Determine required knowledge and skills	What knowledge and skills are required to meet the objectives? Suppose a company wants to increase customer satisfaction. This requires a certain level of product knowledge and communication skills that can be subdivided into written communication skills and/or verbal communication skills and so on. In an academic course on organic chemistry and its significance to carbon dating, it's important to ask what are the knowledge and skills required to comprehend those concepts and properly utilize them. The answers to that question can help the professor establish a topic structure to define the knowledge, skills, abilities, and attitudes required to meet the objectives.
3	Conduct a needs analysis or skills gap survey	This reveals the knowledge and skills people already have as well as what they still need. A gap analysis can be derived from the difference between what is required and people's current knowledge and skills. The analysis might reveal that those taking the assessment have fine verbal communication skills but their written communication skills are poor. That determination becomes critical because the company is moving toward greater use of e-mail communications. The study reveals that the company should establish a course on written communication skills or perhaps create a job aid to help people copy and paste standardized text into e-mails. (Needs analysis often reveals that training isn't the answer.)
4	Develop a learning plan	The plan will describe the learning objectives and explain how the plan will be administered. The learning objectives will guide the production of learning materials and assessments. Facilitating the learning might involve instructor-led training, coaching by managers, or e-learning courses.
5	Run a pre-learning assessment	This a) creates intrigue so that participants are interested in the course, and b) guides each participant to the right learning experience. Some learners will be advanced while others will be novices. The pre-learning assessment guides each individual to an appropriate learning experience.

3.1 The Learning Event

The learning event itself utilizes formative assessments, or forming information into knowledge and providing retrieval practice. A coach or teacher using this technique might say to a learner, “Did you get that?” and then ask the person a specific question requiring the student to remember, consider, and perhaps apply the material. That approach forces search and retrieval practice that helps the learner stand a better chance of remembering the material being taught the next time.

The next step requires a post-course assessment. These measures show whether someone knows enough after the course has been completed, i.e., whether or not the learner has mastered the competencies that the course sought to create. Administering this assessment will determine whether the learner must take the course again, can head off to do his or her work, or should attend a pre-certification exam.

In some instances, particularly in the highest-stakes assessments, organizations will offer pre-certification and practice tests. Companies provide these to prevent any negative reaction to a certification exam. In non-regulatory certifications—Microsoft and others—there will be negative consequences for the company and their product if people keep failing the high-stakes assessments. To head this off they provide pre-certification exams to help get people up-to-speed to pass summative assessments or certification events.

If learners don't do well on a pre-certification exam, they might be looped back again into the course. In addition, after an appropriate learning event, participants might complete a reaction assessment so that teachers and administrators can determine the learners' opinions of the learning experience in order to help them improve the learning experience for others.

3.2 Assessments After Learning

In 1959, Donald Kirkpatrick developed what has become one of the most popular models for evaluating education programs. Kirkpatrick's system has four levels of evaluation, which are as follows:

- **Level 1** measures the reactions of participants after a learning experience. It attempts to answer questions regarding the participants' perceptions: Did they like it? Was the material relevant to their work? Did it meet their expectations? How should the learning experience be improved?

- **Level 2** measures whether or not a participant achieved the learning objectives during the learning event or during a series of such events. It is all very well that the learning experience exceeded the participant's expectations, but if knowledge transfer and competence were not achieved, the validity of the event would be called into question. As an example we might test for written communication skills after instruction to determine if the learner is now qualified for the job.

- **Level 3** measures whether learners were able to apply their new knowledge and skills to their job. We know from Level 2 that they have the skills, but are they using those skills on the job? Are there other issues that are stopping them from being successful on the job? Has their behavior changed? Information for Level 3 evaluations is generally gathered via surveys and personal interviews.

- **Level 4** concerns results. This level tries to answer the question, "Was the learning program effective?" It is good that the learners loved the training, and that they learned everything there was to know and applied everything to their job, but did the expected results appear? Level 4 is a good indicator of whether the learning program had been thought through properly. Was training in fact the issue or would job aids, incentives for good behavior, or consequences for bad behavior have been more appropriate? When done well, Level 4 assessments measure the success of the program in terms that managers and executives can understand: increased production, improved quality, decreased costs, reduced frequency of accidents, increased sales, and even higher profits or return on investment.

There is also another level—Level 5—that Jack Phillips added in the early 1990s. Level 5 measures return on investment. This determines the actual financial benefit to the company against the training investment. This is a challenge to quantify, but the art of data gathering and analysis has progressed dramatically in the last few years and measuring ROI is now practical for larger training programs.

For more information on the four levels of evaluation refer to Evaluating Training Programs: The Four Levels (3rd Edition) by Donald L. Kirkpatrick (ISBN: 10: 1-576753-48-4).

4. Improving Retrieval in Performance Situations

There's a familiar joke about the value of studying:

> *"Why study? The more you study, the more you know. The more you know, the more you forget. The more you forget, the less you know. So, why study?"*

There may be a kernel of truth to this joke, but we can distill a person's ability to retrieve information down to a simple equation:

Retrieval = Learning minus Forgetting

We often focus on climbing the learning curve, but we can also limit our slide down the forgetting curve. Reducing forgetting will improve retrieval.

Dr. Will Thalheimer is an expert on the issue of retrieval in performance situations. He has reviewed numerous studies published in refereed journals and distilled the essence of this research into a series of research-to-practice reports and presentations. Within this paper we'll briefly review some key aspects of Dr. Thalheimer's work:

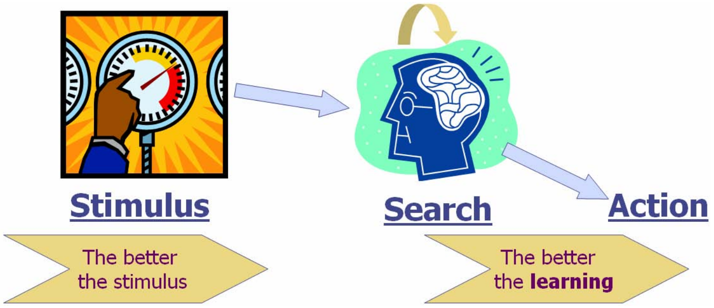
1. Retrieval in a Performance Situation
2. Stimulating Search and Retrieval with Questions
3. Measuring with an Assessment
4. Factors that Influence What People Learn
5. Providing Learners with Feedback
6. The Learning Benefit of Asking Questions

4.1 Retrieval in a Performance Situation

Our ability to retrieve knowledge at the right time is based on many factors, including the stimulus used to trigger our memory search and retrieval. The more similar the stimulus in the learning situation is to the stimulus in the performance situation, the better the ability of the learner to remember what they learned. Questions can act as such stimuli. The more a question can "simulate" the learners' performance situation, the more it will facilitate remembering.

Providing the exact stimulus is sometimes an expensive proposition. However, using a reasonable proxy of the target stimulus can produce substantial learning effects. Let's consider the issue of a call desk agent, whose performance situation includes the audio coming from the phone and the visual information from the computer screen. These are the most important stimuli to practice with during the learning process because these are the stimuli that the learner will have to respond to later when they take real calls. So it would be ideal to use both audio from a phone and visual stimuli from the computer for practice and learning measurement. Multiple choice questions would not do the best job of simulating the performance situation. If audio isn't available, a reasonable substitute is text-based questions that mimic the words of an incoming caller. Using questions to ask merely about features and functions would not measure someone's ability to perform on the job. Ideally, we want to simulate the performance situation as closely as possible.

There are many other stimuli that could be used to improve the ability to retrieve information in a performance situation. For instance, if workers are expected to perform in a noisy factory environment, it might be best if they were trained in that environment. If pilot trainees need to fly in a combat mission, practice in a simulated combat environment would work the best. If you have to perform in a stressful environment, you would do well to learn in that environment.



4.2 Stimulating Search and Retrieval with Questions

We know that repetition helps learning, but constant repetition can be distracting and not very stimulating. We know that repetition helps learning, but constant repetition can be distracting and not very stimulating. We know that repetition helps learning, but constant repetition can be distracting and not very stimulating. We know that repetition helps learning, but constant repetition can be distracting and not very stimulating.

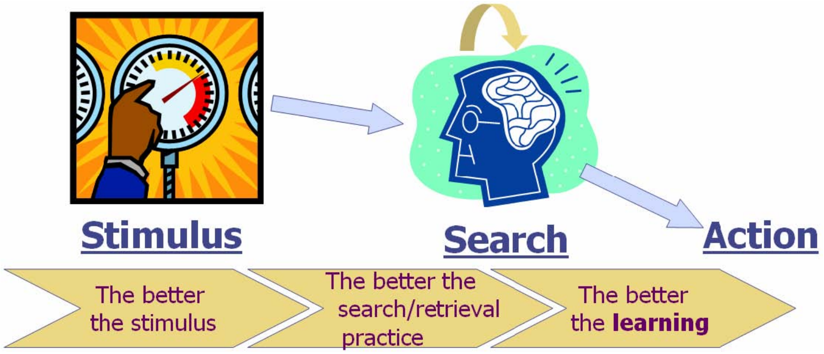
You get the point!

But now, let's ask a question: You want your learners to remember what you're teaching them. What should you do?

1. Avoid repeating the learning material.
2. Repeat the learning material immediately and repeat it exactly as is.
3. Repeat the learning material immediately but paraphrase it.
4. Repeat the learning material after a delay and repeat it exactly as is.

We hope you find this example more stimulating than our repetitive paragraph! Questions provide stimulus and search and retrieval practice, both of which will help us remember in a performance situation. Just as practice helps us master skills, search and retrieval practice helps us remember.

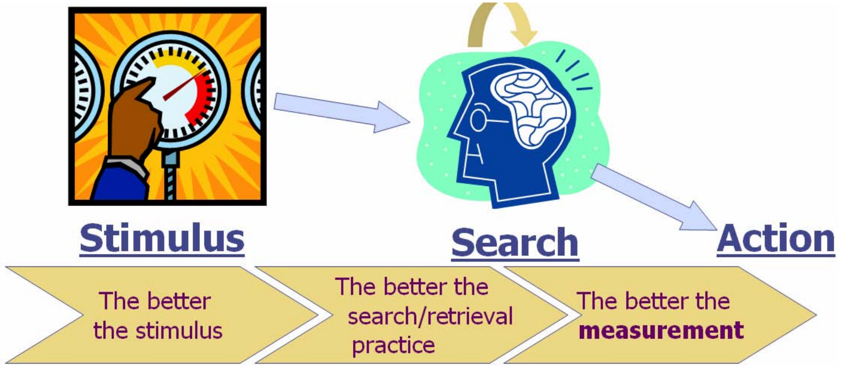
It is best to present questions in an environment that is close to the performance situation. Asking questions provides a stimulus for search and retrieval practice. The stimulus, which could be a voice, text, a multiple choice question or a simulation, is the question itself. The more the stimulus simulates the search and retrieval practice that will be used in the performance situation, the better. The environment for learning is important, too: If someone works in a noisy, dark and damp atmosphere, they should probably learn in a noisy, dark and damp atmosphere, especially if that's where they are expected to do the search and retrieval. If a topic is really difficult for learners, they may need to be initially presented with a simpler--less realistic--scenario. But still, eventually we should move them to as realistic a situation as possible.


Stimulus
Search
Action
The better the stimulus
The better the search/retrieval practice
The better the **learning**

Sometimes high cost can force a compromise in this regard. For instance, it would be extremely risky to put a would-be pilot in a real-life military aircraft to learn how to fly. However, putting that same learner in front of a computer with a multiple choice question doesn't really provide an adequate learning experience either. But if the tests use graphics or video on the computer, that greatly improves the stimulus. Here is another example: In a call center, if instructors add audio of a customer's voice and then require the trainee to reply to the simulated customer's questions, that provides much better stimulus than an ordinary quiz, and therefore, improved learning.

4.3 Measuring with an Assessment

The above principle also applies to measurement: The closer the measurement environment is to the performance environment, the better the measurement of knowledge and skills performed on the job will be. That is why driving tests include actually driving a vehicle. The diagram below expresses this principle:


Stimulus
Search
Action
The better the stimulus
The better the search/retrieval practice
The better the **measurement**

4.4 Factors That Influence What People Learn

Many factors influence how much people learn and what they can retrieve: the learning environment, the manner in which the material is presented, how that material is taught, and others.

In virtual presentations, learners might read their e-mail or talk to someone who comes into their office. Students in a classroom might be looking out the window or thinking about other subjects. Distractions like these can have detrimental effects on learning.

Most people don't absorb everything. They hear it but don't understand what they hear. Sometimes the concepts are too complex for people to initially grasp. They need to hear the information a couple of times before it starts to sink in.

Even if learners absorb everything, there is a good chance they will forget something. Ultimately, we can only remember what we can retrieve. Learners might forget because the stimulus or contextual cues aren't present or because too much time has passed since they learned something.

There are some situations where learners feel that they have learned something but in fact they have misconstrued the information and have developed a misconception.

Assessments play an important part in the learning process. Diagnostic assessments can direct us to suitable learning experiences. Formative assessments can help us enhance learning by directing attention, creating intrigue, providing search and retrieval practice, and correcting misconceptions.

For example, well designed pre-questions can create intrigue before a learning event even starts. And asking questions during the learning event forces students to turn off the distractions and go through that search and retrieval process in order to respond. Questions bring students back on track because they must respond.

But what if we don't absorb everything? ("The more we study…") While it's true that repetition will help learning, if we just keep repeating things we get bored. Asking questions constitutes another form of repetition that doesn't get boring because it forces learners to interact and actually think a problem through.

These techniques also help diminish forgetting. Repetition constantly reinforces new information for learners, and feedback can correct any misconceptions. However, we're often overly optimistic about our ability to remember information. By spacing questions over time, we can reduce the forgetting curve by providing ongoing search and retrieval practice which aids the learning process. If someone had asked you to solve an algebraic equation every week since you left school, the chances are good that you would still be able to solve one. This would be useful if you ever planned to go back to school or use algebra on the job.

Finally, remember that learning is all about the context. Providing retrieval practice in a performance situation helps learners connect their environment with how to retrieve the required information when they need it.

4.5 Providing Learners with Feedback

Giving learners feedback when they take assessments is a valuable way to help them make cognitive connections and help retrieve correct information in the future.

Feedback is particularly useful as a means of dealing with difficult questions, helping participants learn from their errors and setting them on the right track when an error demonstrates a fundamental misunderstanding of a topic.

Stimulus is provided to a participant according to their responses within an assessment. The feedback can be based on how a participant answers a single question or a group or “topic” of questions. It can also be based on how a participant scores on an assessment

When providing feedback, it’s important to give the participant guidance for improving their skill or knowledge. Feedback should be precise enough that the participant knows what needs to be improved, and it should tell the participant how and where to access further learning resources.

Feedback should be used with caution, and in some instances it’s probably best to avoid it. For instance, when security is important, providing feedback may compromise the integrity of the test as a whole. And when questions are related, feedback may end up providing clues about how to answer other questions.

Dr. Thalheimer has examined the latest research on this topic and provides practical recommendations about it in his research paper, Providing Learners with Feedback. The paper emphasizes that there is still a lot more to learn about feedback, but it explains certain principles that should be followed in order to use feedback effectively. Here are a few rules of thumb from that paper, which can be downloaded from http://www.work-learning.com/catalog.

- Feedback works by correcting errors, whether those errors are detected or hidden.
- It should be corrective. Typically, this means that feedback ought to specify what the correct answer is. When learners are still building understanding, however, this could also mean that learners might benefit from additional statements describing the “whys” and “wherefores.”
- Feedback works through two separate mechanisms: (a) supporting learners in correctly understanding concepts, and (b) supporting learners in retrieval.
- To help learners build understanding, feedback should diagnose learners’ incorrect mental models and specifically correct those misconceptions, thereby enabling additional correct retrieval practice opportunities.
- Elaborative feedback may be more beneficial as learners build understanding, whereas brief feedback may be more beneficial as learners practice retrieval.
- Immediate feedback prevents subsequent confusion and limits the likelihood for continued inappropriate retrieval practice.
- Delayed feedback creates a beneficial spacing effect.
- When in doubt about the timing of feedback, you can (a) give immediate feedback and then a subsequent delayed retrieval opportunity, (b) delay feedback slightly, and/or (c) just be sure to give some kind of feedback.

The research indicates that feedback improves learning 15 to 50%. It corrects misconceptions and is best used in response to incorrect answers rather than reinforcing correct answers. It is also good to follow feedback with more retrieval practice. As indicated in the chart below, feedback plays a significant role in helping learners, but retrieval practice is even more crucial.

4.6 The Benefits of Asking Questions

Dr. Will Thalheimer's research of refereed journals reinforces the argument for the learning benefits of questions.

Learning benefits from questions	**Learning Benefit**	
	Min.	Max.
Asking questions to focus attention	5%	40%
Asking questions to provide repetition	30%	110%
Feedback *(to correct misconceptions)*	15%	50%
Asking questions to provide retrieval practice	30%	100%
Questions spaced over time (*to reduce forgetting)*	5%	40%
Potential benefit by using all techniques	85%	340%
Probable learning benefit	42%	170%

Based on research by Dr. Will Thalheimer. Thalheimer (2003). The Learning Benefits of Questions. Available at http://www.work-learning.com/Catalog/

5. Analyzing Results

When reviewing the different forms of assessments, it's clear that they achieve very different ends. Each form of assessment requires a different style of analysis and reporting to reveal the value for the participants and/or the administrator.

Diagnostic assessments, for example, evaluate someone's knowledge, skills, abilities, and/or attitudes and potentially prescribe a suitable learning experience or event. It is useful to determine the various levels of learning events that should be available. These levels can then be used to set diagnostic and potentially prescriptive feedback. But how does one know whether the diagnosis and prescriptions are appropriate? Student surveys and surveys of instructors often reveal the accuracy of diagnostic assessments.

In summative assessments, which are higher-stakes exams, item analysis confirms that the questions are measuring knowledge, skills, abilities and attitudes appropriately and the reports are formatted to support any psychometric review.

We will examine the issues and provide sample reports, taken from real customer data, to help explain how reports can provide the evidence needed to improve the learning experience and derive meaning from the measurements taken during assessments.

5.1 Analyzing Diagnostic Assessments

A diagnostic assessment is primarily used to identify the needs and prior knowledge of a potential student in order to direct him or her to the most appropriate learning experience. There are two aspects of analyzing a diagnostic assessment:

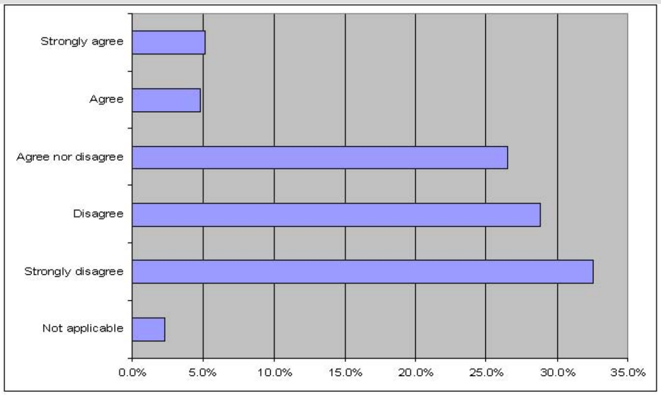
- Initial Analysis: Determines the questions needed and the topics they should cover
- Post Analysis: Determines if the diagnostic assessment is performing to the student's satisfaction.

The initial analysis is conducted using the same techniques as a needs assessment. The post analysis relies on surveying and interviewing participants and instructors and determining if participants were routed to the right kind of learning experiences. The survey results shown below point to a faulty diagnostic assessment, assuming that the Level 1 reactive assessment results showed that the course, environment, course material, and instructor were well received. The results from the Likert scale item, "I felt that the course in which I was placed was too advanced for me," clearly show that something is amiss:

24. I felt that the course in which I was placed was too advanced for me.

Times presented: 78

Times answered: 72



5.2 Analyzing Formative Assessments

A formative assessment's primary role is to provide search and retrieval practice for a learner as well as giving prescriptive feedback (at the item, topic and/or assessment level). In formative assessments, students receive feedback at an item level or at a topic level. This helps the participant understand where they're going right and where they're going wrong. It's not really a report; it's real-time feedback to the learner.

Generally, little analysis is performed on these items; but learners are sometimes surveyed with a reactive assessment (Level 1 survey) to see if the feedback being provided by the formative assessments (quizzes) was useful.

5.3 Analyzing Needs Assessments

A needs assessment is used to determine the knowledge, skills, abilities, and attitudes of a group to assist with gap analysis and courseware development. Gap analysis determines the variance between what learners know and what they are required to know.

There are two key reports for needs analysis:

1. Job task analysis results to reveal actual practice. This is expressed quantitatively.
2. Gap analysis between the abilities required and those demonstrated during a needs analysis skill survey.

The job task analysis (JTA) survey asks questions of subject matter experts and those on the job to determine the significance and frequency of particular tasks. The JTA guides curriculum design and the development of questions to test the knowledge, skills, abilities, and attitudes that relate to a job.

The needs analysis assessment report (below) delivers its scores by topic. When evaluating the two different groups' scores—those of the workers on the job and those of the SMEs—it's clear that knowledge about food safety is the problem. The overall test score does not reveal the issue to address, although it does distinguish a global difference between the two groups. The gap analysis, at a topic level, reveals an actionable theme. Only a topic level score provides the key for a diagnosis.

Topic Name	Average scores from workers	Average scores from SMEs	Gap
Food Safety	60%	93%	33%
Food Packaging	90%	91%	1%
Customer Service	71%	82%	11%
Overall Score	74%	89%	15%

Examining overall results can result in little actionable value. Figure 6 below shows that there is a difference between new employees and SMEs, but it is challenging to determine a suitable course of action:

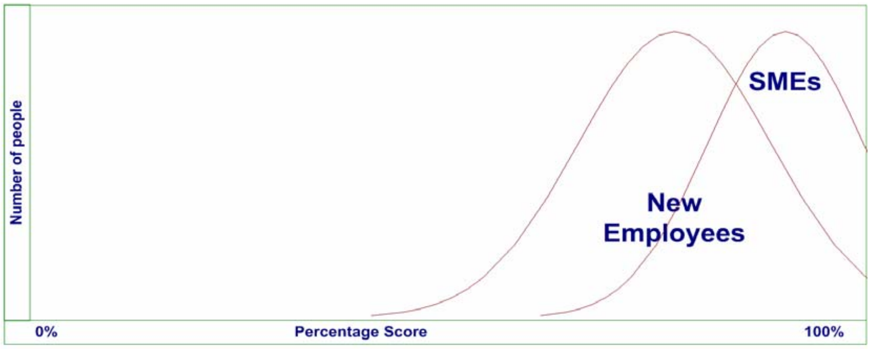


Figure 6 Overall Results

Figure 7 below compares our SMEs to new employees and finds that little training is required to help new employees understand the aspects of packaging and presentation.

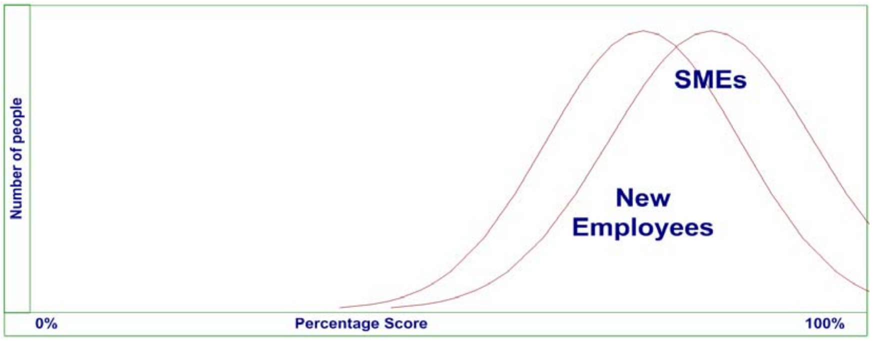


Figure 7 Results on the topic of Packaging and Presentation

Figure 8 below shows the test results at a topic level for "Food Safety." The result of this food safety assessment clearly shows that the new employees have a completely different understanding than the subject matter experts. It helps the employer to understand that the organization needs to invest in new-hire training to ensure that food is prepared safely.

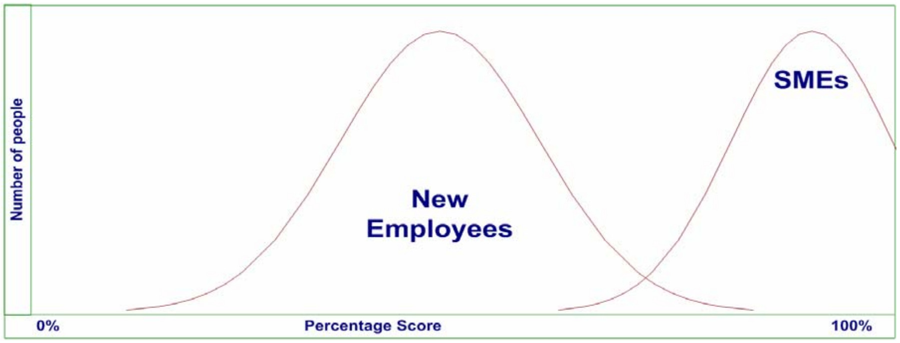


Figure 8 Results on the topic of Food Safety

5.4 Analyzing Reaction Assessments

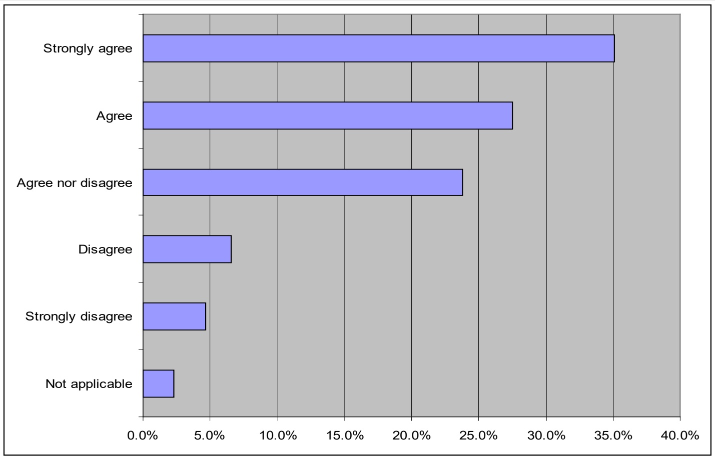
A reaction assessment determines the satisfaction level with a learning event or an assessment experience. These assessments are often known as Level 1 evaluations (as per Dr. Kirkpatrick), course evaluations, or "smile sheets." They are completed at the end of a learning or certification experience.

Reaction assessments aid the planning process for revising a course and its delivery.

In the example below, question 8 and question 9 of this report were interesting:

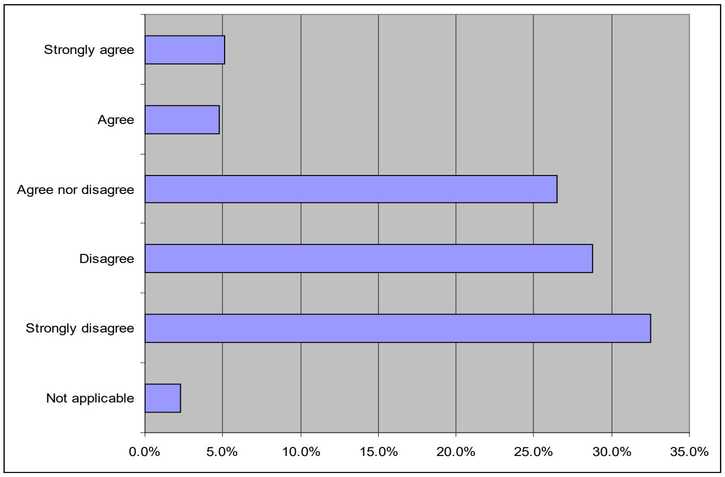
8. I have sufficient time to take the training I need to do my job well.

Times presented: 242
Times answered: 227



9. Training courses I need are scheduled at convenient times and places.

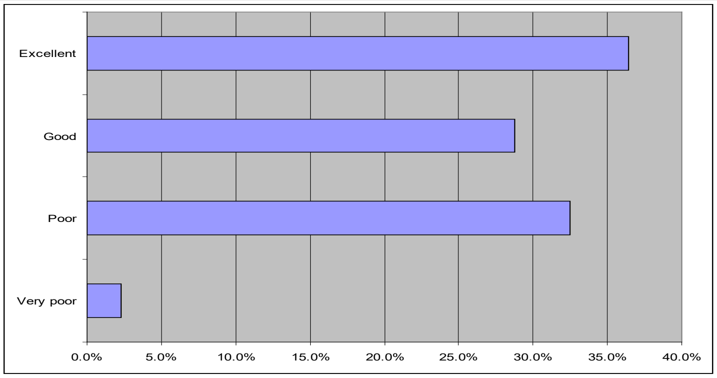
Times presented: 242
Times answered: 227



Eight asks if, “I have sufficient time to take the training I need to do my job well,” and the respondents report that, yes, they have sufficient time. However, all those who responded to question nine disagreed with the statement: “Training courses I need are scheduled at convenient times and places.” What those who administered this assessment discovered was that training events had been scheduled for end-of-the-month periods when the company’s work schedule was quite hectic. Fixing this problem was easy; the company moved the training to the beginning of the following month. The answers to question 9 changed things dramatically.

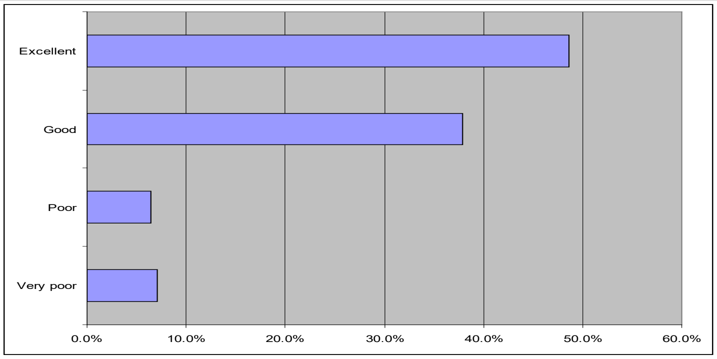
Aggregation of questions about learning environment

Times presented: 239
Times answered: 208



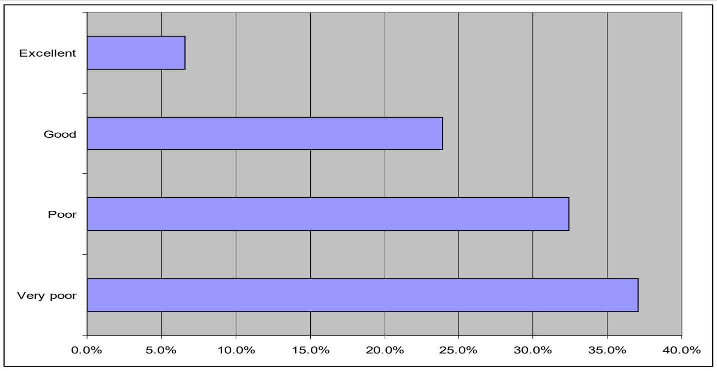
Aggregation of questions about instructor

Times presented: 239
Times answered: 208



Aggregation of questions about the course materials

Times presented: 239
Times answered: 208



The answers to another set of questions—illustrated above—indicate that the training environment and the instructor were wonderful, but the course materials were not. The message is clear: the trainers must revise the course materials.

5.5 Analyzing Summative Assessments

A summative assessment’s primary purpose is to give a quantitative grade and/or make a judgment about a participant's achievement. This is typically known as a certification event if the purpose is to determine whether the examinee meets the predetermined standard for a specialized expertise.

There are several analyses which are required to ensure that a summative assessment is valid and reliable (see below).

5.5.1 Item analysis

The results of this item analysis for a multiple choice question illustrate how tracking individual answers can reveal important information about both the test itself and those who took it.

Choice	# of respondents	% selected	Difficulty	Participant mean	Discrimination	Outcome correlation
A	30	4%		36%		-0.32
B	75	10%		47%		-0.22
C	480	64%	.64	76%	0.49	0.24
D	165	22%		52%		-0.15

Among the choices in the chart above, C was the correct answer (highlighted in green). As the chart illustrates, four percent chose A, 10 percent chose B, and so on. What do these statistics reveal?

First of all, if 100 percent of those taking the test got the correct answer, it would mean that the question was too easy to help separate the performances of the test takers from one another. If you are creating a norm-referenced test designed to select only the best performers from within a group of test takers, items that all examinees get correct are not helpful. If no one got the correct answer, it would mean that the question was too difficult to separate test takers' scores from one another, and again, is not very useful for creating a norm-referenced test. Questions that distinguish maximally between knowledgeable candidates and not so knowledgeable candidates are the useful questions when your goal is identifying either the best or poorest performers within a group.

Criterion-referenced test writers have a somewhat different attitude toward test questions that all test takers get correct or incorrect. This kind of test is composed of items that measure specific and important competencies that test takers must have in order to do a job or succeed in a course. Items that measure such competencies in the eyes of SMEs are included on the criterion-referenced test even if all test takers generally get them correct or even if few test takers get them right. The point is that the achievement of each competency must be verified by the test taker, and examinee scores are compared to a predetermined standard rather than compared to one another. If instruction in important competencies is successful, there may be many items that all test takers get correct. Because the goal of the test is not to spread the test takers scores out from one another, such items present no problems unless examinees are only getting them correct because the questions contain clues to the correct answer rather than measuring genuine competence. Items that all test takers miss indicate serious problems with either the items themselves or with the instruction that prepares examinees to take the test. If such items are not flawed, they remain on the criterion-referenced test, and the instruction is improved until test takers can answer them correctly.

The "Difficulty" index is the percentage of examinees who answered the item correctly. In a typical four-choice multiple choice question, if all the test takers guessed, approximately 25 percent would select each choice. As a result .25 is normally the lowest difficulty number you'd expect to see if none of the test takers know the answer because the item is flawed or the preparatory instruction is bad. Typically, values for difficulty will be in the range of .6 to .8, but these percentages depend on the quality of the question and the competence of the test takers.

If test creators want to spread out the scores of examinees from one another, it is necessary for them to include more relatively difficult questions and questions that discriminate between high performers and low performers on the test as a whole. To simply make a test more difficult to pass, it isn't necessary to start thinking up more difficult questions. Begin by ensuring that the questions are valid and provide the right focus on important objectives and then simply adjust the pass/fail score upward.

The column titled the "Participant Mean" presents the average final score on the entire assessment of those examinees that chose the corresponding optional answer. The people who selected choice A averaged 36% on the entire test, illustrating that less knowledgeable students on the overall test were also choosing the wrong choice on this question. On the other hand, the students selecting the correct choice, C, were in fact the more knowledgeable students with a participant mean of 76 percent.

Discrimination is a number that ranges from -1.0 to + 1.0 and illustrates how the results on this question compare to results on the test overall. Looking at a specific example will make this clearer. If we imagine nurses taking an exam, and the more knowledgeable nurses get a question right while the less knowledgeable nurses get the question wrong, that will result in a higher "discrimination" – the number in the sixth column. In our example the discrimination is +0.49.

Now imagine that a question about baseball is inserted into a nursing test. It's very possible that a lot of nurses know a great deal about baseball, but it is unlikely that the more knowledgeable students of nursing will also know more about baseball. The test scores might illustrate that some nurses are very knowledgeable about nursing and very knowledgeable about baseball while others who are very knowledgeable about nursing aren't knowledgeable about baseball at all. So it is highly unlikely that the results from this question on baseball would correlate to the overall test results. It is highly likely that the results would reveal a correlation near zero (sometimes called a "zero order correlation") because this particular question on baseball has no relationship to the overall test results.

Another way of looking at this is that we are now measuring two things: knowledge of nursing and knowledge of baseball. How much nurses know about baseball has no relevance to their nursing skills: there is no reason to believe that knowledge of one is related to knowledge of the other. So it is highly unlikely that the results from this question on baseball would correlate to the overall test results. The Discrimination and Correlation statistics help us identify these issues.

The last column in the chart shows a statistic known as Outcome Correlation. This question performed well because the wrong choices have negative correlations (the students choosing them didn't do well on the test overall) and the right choice has a positive correlation, which indicates that the item discriminates between good and poor performers on the test. High discrimination indices are particularly important to norm-referenced test creators who seek to spread out the test scores. Criterion-referenced test designers are alert primarily to items with negative Outcome Correlations for the correct choice. These items are almost always flawed because high test performers consistently miss them while poor performers get them correct. These items should be reviewed and revised or discarded.

5.5.2 Setting Pass/Fail Score for Norm-Referenced Tests

As discussed earlier, norm-referenced tests are designed more for ranking and comparing how a test taker compares against a group – often to aid in a selection process – rather than achieving a certain minimum "passing" score. If there is a "pass/fail" score, it is generally based on a predetermined number of candidates/examinees that are expected to pass. A report shows how many people reached each score, enabling test administrators to select the top set of candidates. For example, using the table below, if 1,000 students should pass for graduation to the next level of their studies, a passing score of 78.6% would achieve that result.

Scores	Number of candidates
0% and above	1,500
77% or above	1,318
78% or above	1,214
78.1% or above	1,156
78.2% or above	1,034
78.3% or above	1,028
78.4% or above	1,015
78.5% or above	1,004
78.6% or above	1,000
78.7% or above	998
78.8% or above	993
78.9% or above	982
79% or above	961

5.5.3 Setting Pass/Fail Score for Criterion-Referenced tests

This paper does not attempt to address all of the issues related to setting a pass/fail score for criterion-referenced tests. However, it does illustrate some reports that are useful to provide evidence for the decision making process.

One method of gaining evidence is to compare learners who are assumed not to not know as much about a subject to those who demonstrate the skill levels the test designer wants to measure. For example, an internal certification program for repair technicians might be based on two groups of test takers: those whom managers identify as having the desired skill levels (skilled-level SME incumbents) versus those who do not (non-SME incumbents). The skilled-level SME might be defined by measures such as repairs that are fixed the first time with no more than a 10% call-back, repairs that are made independently without need to call the help desk or repair rates that meet operational goals such as six visits per day. Figure 9 illustrates how using both types of incumbents for establishing the cut score can provide easiest strategy for most organizations.

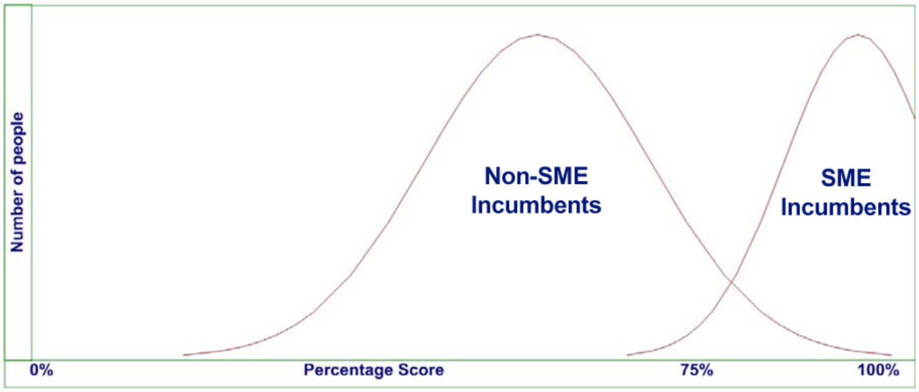


Figure 9 Comparing scores of skilled and non-skilled job incumbents

If we look at the overall results, we see how non-SME incumbents performed compared to the scores of SME incumbents. For non-SME incumbents, the bell curve is toward the lower end of the score scale, while for SME incumbents the scores move to the higher end.

One technique for setting the pass/fail is to choose the score where these two curves intersect. In the example above that would be at about 80%. The score at the intersection minimizes the number of test takers in these groups that were misclassified by the test, i.e., minimizes the number of non-experts who passed and the number of SMEs who failed.

If the pass/fail decision is determined only by a single score, then a single set of results can be used as evidence. However, if multiple cut score levels are used for the various topics in the test then multiple sets of scores must be examined.

6. How Technology Can Improve the Assessment Process

Clearly, assessments of all types can have a major impact upon what and how people learn, whether it's in a corporate environment or on a college campus. It's also true that if these same organizations had to deliver these kinds of assessments on paper – the formative, needs, reactive, summative – it would be very time consuming and difficult to gather the results needed to effect change.

Technology makes it possible to create assessments quickly and cost-effectively deliver them online, offline and even on paper, then process and report on the results centrally.

Web-based authoring tools expedite the production of assessments by giving subject matter experts an easy way to author and share questions. Auto-sensing and auto-sizing capabilities make it possible to deliver a single assessment to many different types of environments, from mobile devices to wikis, blogs to portals, thereby making assessments an integral part of the learning process. Item banking provides effective storage and categorization of questions and can be drawn upon to create individualized assessments. Translation management tools help multinational organizations keep track of multiple versions of the same questions and provide localized content. Efficient integrations with enterprise systems make it possible to administer assessments seamlessly and harvest the results for future use.
The benefits of assessment technology to learners are immense: Questions can incorporate photos, graphs and other visual aids, as well as multimedia files. They can provide an interactive test-taking experience that can simulate the participants' work environment. Tools such as calculators can be made instantly available within a test. Rapid feedback informs participants of areas that require more focus and can direct them to relevant learning materials. The ability to take low-stakes assessments anytime, anywhere, also gives learners increased opportunities for retrieval practice.

Technology enhances the assessment process in many other ways:

- Security
  - Secure browsers can lock down desktops to prevent task switching
  - Monitors can be required to log in along with a test taker
  - Role-based access protects content and results
- Accessibility
  - Text-sizing and contrast controls, navigation options, keyboard shortcuts and audio files help to accommodate various disabilities
  - Assessment delivery platforms can be optimized to work with such assistive technologies as screen readers and alternative pointing devices
- Authoring
  - Browser-based authoring tools mean subject matter experts can contribute content quickly and easily
  - The elimination of word processing documents and emails enables a more efficient workflow and collaboration
  - Enabling collaborative authoring in the same building or around the world. Subject matter experts can contribute new questions, share question sets and agree on final versions of questions
  - Secure, globally accessible item banks and repositories can be accessed by authors and administrators worldwide
- Item banking
  - Once items are created, they can be used in multiple assessments, thus making more use of each item, with a centralized results set for reporting analysis
  - Items can be updated centrally, making it easy to keep assessments up to date
- Integrations with enterprise systems
  - Assessments can be accessed from within other systems
  - Results can be passed back for archiving along with other records
- Translation management
  - Online tools make it possible to maintain various versions of a single item or assessment
  - Localization of content is more efficient, with a smoother workflow
  - Participants can take assessments in languages that meet their needs

    - Results from multiple locations can be consolidated into a single results set for reporting
- Centralized record keeping
    - Aids organizations that need to demonstrate compliance
    - Provides an audit trail
- Blended delivery
    - Maximizes the number and types of devices on which to deliver assessments
    - Provides options for delivering assessments to people with intermittent internet connections
    - Means inexpensive handheld devices can be used to deliver assessments
    - A single assessment can be delivered via computers, hand-held devices, or on paper, with all the results being processed together
    - Enables embedded assessments within learning materials blogs, wikis and portals, thereby providing retrieval practice along with learning

7. Conclusion

To yield real benefits from the assessment techniques described in this paper, begin by identifying your goals: Do you need to identify qualified people, improve customer service, improve response times, or meet regulatory compliance?

Next, document the topics and learning objectives. Determine the style of assessments that your organization needs to achieve the goals you have set out. These assessments will enable your organization and your learners to reach their (and your) objectives.

Finally, decide how you'd like to administer these assessments.

Questionmark can provide technologies and services to help you achieve the results you seek. You can learn more from www.questionmark.com.

Recommended Reading

Criterion Referenced Test Development: Technical and Legal Guidelines for Corporate Training and Certification by Sharon A. Shrock and William C. Coscarelli (ISBN 10: 0787988502)

Evaluating Training Programs: The Four Levels by Donald L. Kirkpatrick (ISBN: 10: 1-576753-48-4)

Tests That Work by Odin Westgaard (ISBN 0-7879-4596-X)

Topics in Measurement: Reliability and Validity by W. Dick and N. Hagerty (ISBN 0070167834)

Work-Learning Research white papers by Will Thalheimer (at www.work-learning.com)
The Learning Benefits of Questions
Measuring Learning Results
Providing Feedback to Learners

White Papers available from the Questionmark Web site:
http://www.questionmark.com/whitepapers

Glossary

Assessment	Any systematic method of obtaining evidence from posing questions to draw inferences about the knowledge, skills, attitudes, and other characteristics of people for a specific purpose.
Exam	A summative assessment used to measure a person's knowledge or skills for the purpose of documenting their current level of knowledge or skill.
Test	A diagnostic assessment to measure a person's knowledge or skills for the purpose of informing the person or their teacher about their current level of knowledge or skill.
Quiz	A formative assessment used to measure a student's knowledge or skills for the purpose of providing feedback to inform the student of their current level of knowledge or skill.
Survey	A diagnostic or reaction assessment to measure the knowledge, skills, and/or attitudes of a group for the purpose of determining the needs required to fulfill a defined purpose.
Diagnostic	An assessment that is primarily used to identify the needs and prior knowledge of participants for the purpose of directing them to the most appropriate learning experience.
Formative	An assessment that has a primary objective of providing search and retrieval practice for a learner and to provide prescriptive feedback (item, topic and/or assessment level).
Likert scale	A method to prompt a respondent to express their opinion on a statement being presented. Likert scales are often 4 point scales (strongly agree, agree, disagree, strongly disagree), 5 point scales (strongly agree, agree, neutral, disagree, strongly disagree), but sometimes present as many as 10 potential choices.
Needs	An assessment used to determine the knowledge, skills, abilities, and attitudes of a group to assist with gap analysis and courseware development. Gap analysis determines the variance between what a person knows and what they are required to know.
Reaction	An assessment used to determine the satisfaction level with a learning or assessment experience. These assessments are often known as Level 1 evaluations (as per Dr. Kirkpatrick), course evaluations, smile or happy sheets. They are completed at the end of a learning or certification experience.
Summative	An assessment where the primary purpose is to give a quantitative grading and make a judgment about the participant's achievement. This is typically known as a certification event if the goal is to document that the test taker has specialized expertise.
SME	Subject Matter Expert

About Questionmark

Questionmark assessment and portal solutions enable organizations to measure knowledge, skills and attitudes for certification, channel expertise, workforce learning and regulatory compliance. Questionmark's assessment management system, available as a cloud-based solution or for on-premise deployment, enables collaborative, multilingual authoring; multiple delivery options including mobile devices; trustable results and comprehensive analytics.

Complete details are available at https://www.questionmark.com

Question*mark*
35 Nutmeg Drive
Trumbull, CT 06611
USA
Tel: (800) 863-3950
Fax: (800) 339-3944
info@questionmark.com

Question*mark*
30 Coleman Street
London EC2R 5AL
United Kingdom
Tel: +44 (0)20 7263 7575
Fax: +44 (0)20 7263 7555
info@questionmark.com

Legal note

Although Questionmark has used all reasonable care in writing this document, Questionmark makes no representations about the suitability of the information contained in this and related documents for any purpose. The document may include technical inaccuracies or typographical errors, and changes may be periodically made to the document or to the software referenced. This document is provided "as is" without warranty of any kind. See your Perception support contract for further information.

Company and product names are trademarks of their respective owners. Mention of these companies in this document does not imply any warranty by these companies or approval by them of this guide or its recommendations.