# Assessments Through The Learning Process

This paper explores how instructors and organizations can use assessments to improve the learning process. It is designed to help readers distinguish between different types and styles of assessments, understand various assessment tools, and learn how to develop effective assessments, analyze their results and appreciate the benefits of computerized assessments.

Authors: Eric Shepherd
Janet Godwin

 *© Copyright Questionmark Corporation 2004* 

The shift in the way that training and classroom courseware are delivered has been profound during the past five years. While a great deal of attention has been given to form of delivery, it's clear that search and retrieval practice as well as other important factors influence how and what students actually learn and apply to their specific tasks.

Some organizations have tried to move their entire organizations to online learning opportunities, for example, and yet not everyone is well suited to this approach. A significant portion of the population, especially those considered to be Gen X or Gen Y, love to learn on the computer. They even spend their leisure time on the computer. Others learn better in the more traditional training session where people get together, break up into groups, and talk to each other. Many of us prefer the classroom setting with all of its human interaction.

It's critical, therefore, to understand how people learn, what they have in fact learned, and whether this knowledge is useful for their particular job. That's why the cornerstone of developing successful educational, training, and certification materials is the effective use of assessments. While assessments used to consist of reams of paper upon which students filled in tiny boxes, now companies and educational institutions have a real opportunity to use technology not only to make assessments more widely available but also to make the process far more effective.

These organizations can use assessments to 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 real 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 the types and styles of assessments
- Distinguish 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


*© Copyright Questionmark Corporation 2004*

## 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 students' knowledge, skills, abilities, and attitudes.

The table below further defines these terms:

| | |
|---|---|
| **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 student's knowledge or skills for the purpose of documenting their current level of knowledge or skill. |
| **Test** | A diagnostic assessment to measure a student's knowledge or skills for the purpose of informing the student or their teacher on 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 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 search and retrieval practice for a student 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 student 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 |


*© Copyright Questionmark Corporation 2004*

Level 1 evaluations (as per Dr. Donald Kirkpatrick), course evaluations, smile or happy sheets and 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.

### 1.1.1 Diagnostic Assessments

If you go to the doctor and just say, "I have a pain," and the doctor says, "Oh great, here's a pill," you'd be concerned. However, what the doctor actually does is ask, "Where is the pain? How often does the pain come? Have you done anything recently to cause the pain?" These are questions the doctor asks to tease out the issues so that he or she can make a diagnosis and write a prescription. That's exactly what happens with 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 for whom English is a second language might take a test to discover if their English skills are adequate for taking other courses. The test measures that person's current knowledge and skill to provide feedback to the instructor so that they can tailor the course effectively. These kinds of tests also create intrigue and, by so doing, will actually increase the learning benefits of the learning experience. For instance, if an instructor asks a question that a student doesn't know the answer to, that student might become curious to find out the answer and, therefore, pay more attention in the class.

Therefore, diagnostic assessments are used to determine knowledge and identify skills gaps and needs. Such an assessment might report that a learner knows everything there is to know about Microsoft Word but only knows 50 percent about 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 for and retrieve the information. If a person gets the question wrong, 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." 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 a student that they're actually learning or 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 learn their level of competence but they also inadvertently reduced their forgetting curve by experiencing some search


© Copyright Questionmark Corporation 2004

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 rather like a series of small tests that provide evidence for the instructor to make a judgment.

### 1.1.3 Needs Assessments

Needs assessments will assess knowledge, skills, abilities and attitudes of a group to help someone determine the training needs of a group or provide data for a job task analysis. These are low stakes assessments that measure against requirements to determine a gap that needs to be fulfilled. 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

Reaction assessment occurs when we assess the reactions and opinions of a student 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 one assessment. In colleges and universities it's called a course evaluation. Such an assessment gathers opinions from the student about what they thought of the course material, of the instructor, the learning environment, of the directions to the facility, and of the audio-visual equipment. From that information the instructor can improve the learning experiences in the future.

### 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. Regulatory and non-regulatory exams, which provide a quantitative score signifying how much knowledge or skill someone has, are an example of a summative assessment.

## 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 not only is there more than one kind of assessment but they can be categorized in terms of the stakes involved in taking the test. These stakes are identified as:

- High
- Medium
- Low

The level an assessment's stakes refer 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 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 wouldn't normally check the ID or watch someone taking a low stakes assessment whereas with a high stakes exam they would.


*© Copyright Questionmark Corporation 2004*

The requirement for validity and reliability is low for a quiz or a survey but a high stakes test must be valid and completely reliable. These higher stakes assessments require more planning. The general rule for a test or exam is that in a work setting they should look like the job, or at an academic institution the test or exam should look like the curriculum.

Very little planning goes on with regard to a low stakes assessment. Subject matter experts (SMEs) simply write the questions and make them available to students. However, a high stakes test requires a great deal of planning such as job task analysis, setting the pass/fail scores, specifying the methods and consistency of delivery required, and 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 are they to the job. Test developers plan which questions should be in the test by topic, which subjects are more important, which have less importance and the depth of competency required. The pass/fail score or cut score determines the threshold between passing and failing.

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 one, however, it's rare to involve psychometricians.

## 1.3 The Stakes of an Assessments

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 | Maybe important | Very important |
| **Proctoring required** | No | Sometimes | Always and constant |
| **Development effort** | Minor | Medium | Major |
| **Check reliability and validity** | Rarely | SME | Psychometrician |

As you can see in the chart above, the consequences different types of assessments vary. A high stakes exam might determine whether a person might be hired or fired or 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


© Copyright Questionmark Corporation 2004

certification exam appealing it would be very difficult, if not impossible. The options for the participant vary in direct relation to the stakes.

Obviously on a survey there's no motivation to cheat and on a low stakes quiz very little. 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 so there certainly is a motivation to cheat. As a direct consequence it becomes more important to identify each test taker. In fact, for high stakes tests related to national security such as for the CIA or the military these organizations might use biometric screening such as retinal scans to ensure that someone is who they say they are.

Obviously if there's a low propensity to cheat there's no need to proctor an assessment, but if there's a high propensity to cheat or high motivation to cheat then there should be constant and continuous proctoring.

The development effort for a quiz is quite minor. Not so for medium and high stakes assessments. A rule-of-thumb for a medium stakes test is that the average subject matter expert will develop three questions an hour, and of those, one will get into the test. High stakes questions take far more time to develop. The average cost of a high stakes question, by the time an SME develops a question and it gets into a high stakes exam, will is between $500 and $1,500 – each! It's important, when the stakes are high, 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 and 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 The Nature of Assessments

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 |

For instance, placement tests constitute a popular example of a diagnostic assessment. People take this sort of assessment so they'll be placed into the right learning experience. Just as with other assessments some placement ones are low stakes and some have higher stakes. For instance, a college entrance exam is a higher stakes placement test but a guide to knowledge and learning resources is low stakes.

If a company has 100 employees and wants to find out what the skill gaps are in order to offer the right training programs for these people, they are performing a low stakes assessment. There are no consequences to the individual candidate. There might be some consequences to the organization, but generally even these are considered low stakes. The same applies to self-assessments. These help people see how they are doing. Some companies use a Monday morning wake up quiz for help desk employees. They've been gone for the weekend and are not really alert. A quiz on the Monday morning gets them into the thinking mode. It's a low stakes formative assessment that provides them with search and retrieval practice.


© Copyright Questionmark Corporation 2004

Medium stakes assessments, on the other hand, measure the level of employees' knowledge and skills. The more personal these get the higher stakes they become, with greater consequences. Medium stakes exams will have consequences because some people are probably going to get paid better for handling more difficult problems. In an academic setting, instructors use this level of assessment to assign grades.

High stakes exams consist of regulatory certifications for groups such as plumbers, electricians, policemen, therapists, doctors, or nurses. These professions are all regulated by city, state, or federal government, so they're considered to be high stakes. However, non-regulatory exams such as the Microsoft, Cisco, or Linux certification exams are also high stakes. Because there's no immediate consequence, they're slightly lower stakes than regulatory certifications, but as they can provide employment and promotional opportunities and therefore consequences to a candidate they have high stakes. Similarly an entrance or pre-employment exam has high stakes. Obviously if an applicant fails his or her entrance exam to law school, the stakes have been high.

Finally, some companies do what's known as granting permission. If personnel pass a exam they can operate a particular kind of machinery. Generally the more dangerous the machinery, the higher the stakes; the less dangerous the machinery, the lower the stakes.

## 1.5 Consumers have a stake in assessments

When a plumber comes to your house are they going to end up breaking a pipe? Is a surgeon qualified to take out someone's gall bladder? Is someone qualified to drive a passenger vehicle safely? While someone might feel sorry for an unfortunate student who fails a test, the more important concern is whether people can be trusted to perform a particular task or job.

As a result, a kind of partnership has developed among consumers, those who take assessments and the people who create the tests. The consumer wants to know that they can trust people they hire;the designer of a high stakes test wants to measure accurately and have a reliable and valid test; and candidates want the tests to be fair.

There needs to 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 referred to as Face Validly. Not only does a test need to be valid from a content perspective so that it fairly represents the tasks of a job, but it needs to be trusted by the consumer and the candidate taking the assessment. The test developer needs to be trained properly; the student needs to be educated about the value of an assessment and be reassured that they're ready for the experience; and the consumer needs to be educated about the validity of the assessments so they can trust the people doing their work.


*© Copyright Questionmark Corporation 2004*

## 2. The Reliability and Validity of Assessments

An assessment is reliable when it works consistently. 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.

If an assessment is valid, it looks like the job. To assess this validity the person 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. From that information it becomes possible to produce a valid test.

An example illustrates how the validity of an assessment can be assured. If a test is administered to a group of skilled nurses and some nurses do very well and some nurses do very badly, then it's clear that the test is not reliable. If it's not reliable, it cannot be valid; these were all good nurses and inconsistent and unreliable results came back. The test is not valid because it doesn't measure the aspects of the job.

**Figure 1**
Reliable (Consistent) but not Valid

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

**Figure 3**
Reliable and Valid

Concept reproduced with the kind permission of William Coscarelli & Sharon Shrock of Southern Illinois University at Carbondale, authors of the book, "Criterion Reference Test Development: Technical and Legal Guidelines for Corporate Training and Certification."

In the figures above the segments of a dart board represent the domains of knowledge or skills required for different jobs. The domain that we need for the job we are testing for is represented by the bull's eye.

The dart board in Figure 1 above shows that all the darts are stuck in the same area, illustrating that a particular assessment is reliable and consistent, but unfortunately it's not valid. If it was valid, all the darts would be in the center. In Figure 2 the darts have landed all over the board. This assessment cannot be reliable because it's not consistent. Finally, the last example is reliable because all of the scores are clustered together; even more importantly, they're all on target with the job.

### 2.1 Assessment Score Interpretations

When people take an assessment it's important for them to understand the implications of their score, particularly when passing or failing make a major difference in their lives. There are two ways to score an assessment. These are referred to as criterion referenced and norm referenced.


© Copyright Questionmark Corporation 2004

With a criterion referenced score interpretation, the test designers have established 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.

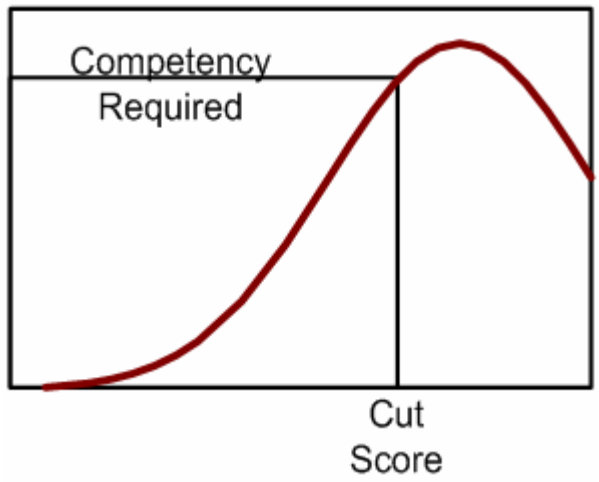



**Figure 4**
Typical bell curve for a criterion referenced test

This bell 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 competency required to pass the exam.

With a criterion referenced score interpretation more or fewer people will qualify from examination event to examination event, since each sitting will yield candidates with more or less knowledge. What's important, however, is that a benchmark has been established for the standards required for 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 those passing against the population's norm. For example, a college entrance exam might be designed to fill 100 spaces in a college. The test determines the 100 best people of those taking the test to fill those positions. Some years a higher quality group of students will qualify and sometimes a lower quality group. The key, however, is that the pass/fail score is set against a population.


*© Copyright Questionmark Corporation 2004*

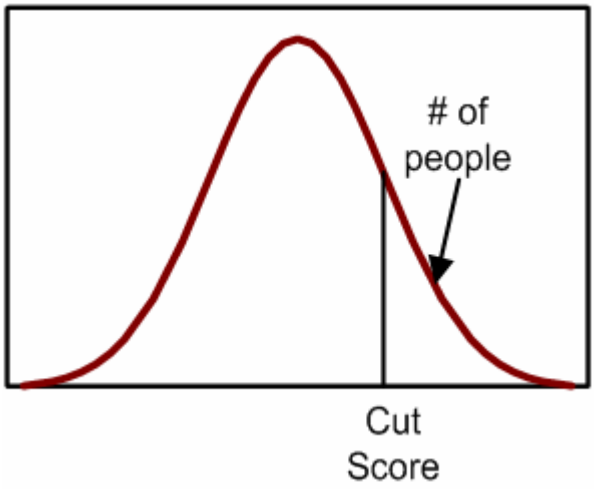



**Figure 5**
Typical bell curve for a norm referenced test

How are these references 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 versus Speeded Assessments

Most tests are timed, but studies have shown that 95 percent of the students will typically complete an exam within the time limit, providing it's been set reasonably. However, there are tests that have to be speeded because speed is an important part of on-the-job performance. A test for a nuclear reactor room technician provides a good example. The test might simulate a dangerous situation in which the person is required 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 which button to push and do so within a given time limit. That is a speeded test because speed and the reaction time is a key part of the actual test and the job.


© Copyright Questionmark Corporation 2004

## 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 six-step process.

First, organizations assess their objectives. The objective might be to increase customer satisfaction or to reduce error rates or improve the safety record of a factory. In a college environment, one would need to ask, "What are the objectives of this curriculum?"

Second, it's important to ask what knowledge and skills are required to meet those objectives. In a business environment, a company may want to increase customer satisfaction. This requires a certain level of product knowledge and communication skills. These can then be subdivided into written communication skills and/or verbal communication skills and so on. Similarly, in a college or university course on organic chemistry and its significance to carbon dating, it's important to ask what are the knowledge and skills required to understand those concepts. From the answers to that question the professor can establish the topic structure to define the knowledge, skills, abilities, and attitudes required to meet the objectives.

The next step requires a needs analysis survey or skills gap survey. Here, people take a needs assessment to reveal the knowledge and skills they already have as well as what they still need. A gap analysis can be derived from the difference between what is required and what is available. That analysis might reveal, for example, that all of the verbal communication skills of those taking the assessment are fine, but the written communication skills are poor. That determination becomes critical because the company is moving toward greater use of e-mail communications. Therefore, the study reveals that the company should establish a course on written communication skills.

Needs analysis often reveals that training isn't the answer. For example if people have to repeatedly send a long, complex response to customers for a common request the company might create a job aid to help people copy and paste standardized text into e-mails.

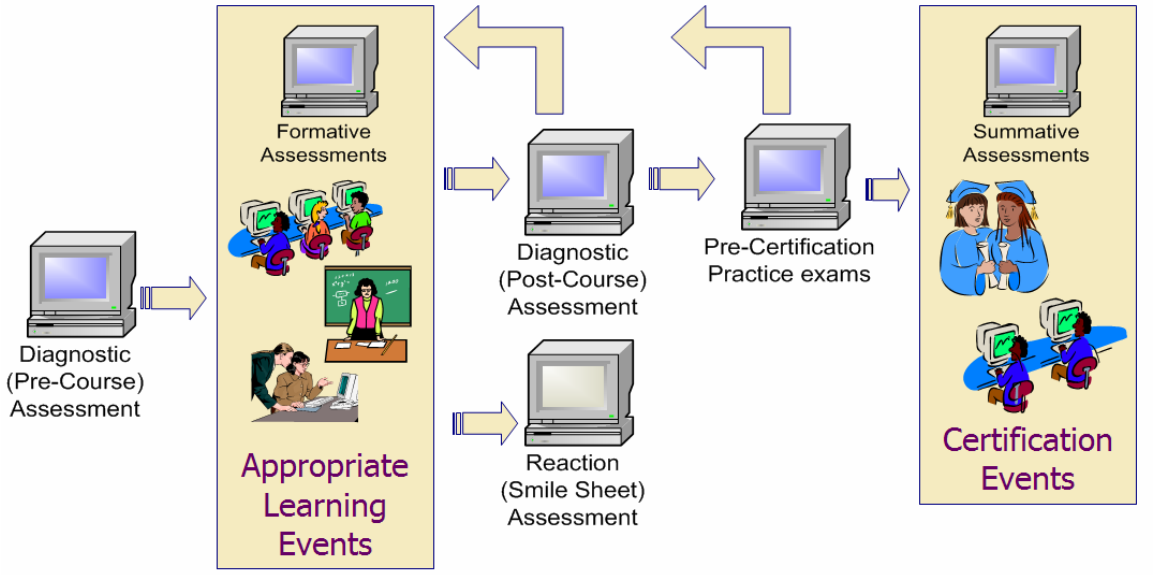
Following these steps allows organizations to develop a learning plan – the fourth step. That plan will include the learning objectives as well as how it will be administered. The learning objectives will guide the production of learning materials and assessments. Facilitating the learning might involve instructor-led training, coaching by managers, or e-learning courses.

Next it's important, before anyone comes into the course, to conduct a pre-learning assessment. The pre-learning assessment will have two purposes: to create intrigue so that students stay awake during the course and, secondly, to guide each student to the right learning experience. Some students will be advanced while others are novices. The pre-learning assessment guides the person 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 student, "Did you get that?" and then ask a question about the material. That approach forces search and retrieval practice that helps the student stand a better chance of remembering the material being taught the next time.


*© Copyright Questionmark Corporation 2004*




The next step requires a post-course assessment. These show whether someone knows enough after the course has been completed. Administering this assessment will mandate whether the student must take the course again, can head off to do their work, or 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 students 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, students might also complete a reaction assessment so that teachers and administrators can find out about their opinions of the learning experience 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.

> **Level 1** measures the reaction 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 how much a student learned during the event or during a series of events. It is all very well that the learning experience exceeded the participant's expectations, but if there wasn't any knowledge transfer the validity of the event would be called into question. As an example we might test for written communication skills after an event to determine if the person is now qualified for the job.


© Copyright Questionmark Corporation 2004

**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 tried to answer the question, "Was the learning program effective?" It's fine that the person loved the training, 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? This measures 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 '90s. 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* by Donald L. Kirkpatrick (ISBN: 1-576750-42-6).


*© Copyright Questionmark Corporation 2004*

## 4. Improving Retrieval in Performance Situations

There's a joke that's been repeated by students for years concerning 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?"

Well maybe there's a kernel of truth to this joke. More realistically we can distill a person's ability to retrieve information down to a common sense equation:

**Retrieval = Learning – Forgetting**

We often focus on climbing the learning curve, but we could also benefit from limiting our slide down the forgetting curve. Simply put, reducing forgetting will improve retrieval.

Dr. Will Thalheimer is one of the few experts on the issues 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 white papers 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. The learning benefits of questions

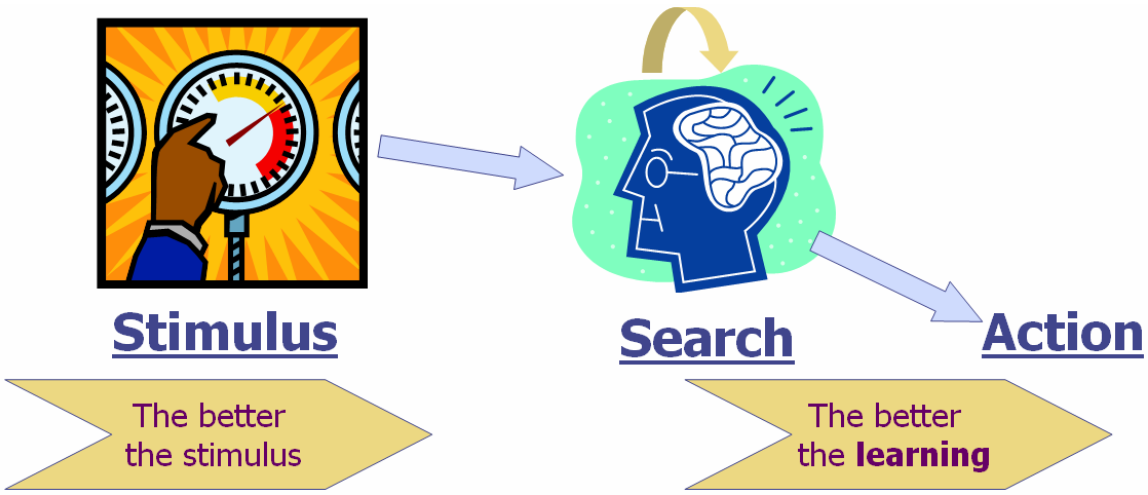
### 4.1 Retrieval in a Performance Situation

Our ability to retrieve knowledge at the right time is based on many factors. One key factor is that the stimulus used to trigger our knowledge search and retrieval process. The closer the stimulus can be between the learning and performance situations the better our ability to remember in the performance situation.

Providing the exact stimulus is sometimes an expensive proposition. However, good-enough stimulus is better than ignoring the issue. Let's consider the issue of a call desk agent. Most of the stimulus will be via audio from the phone, but the agent also gets visual information from their computer screen. When educating them and measuring their ability to perform it would be better to provide audio rather than text stimulus and better to provide screen displays to which they can respond. Multiple choice questions would not do such a good job of simulating the performance situation. If audio isn't available it would be best to use text that best represents the situation of an incoming caller. Using questions to ask merely about features and functions would not measure the ability of someone to perform on the job.

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.


*© Copyright Questionmark Corporation 2004*




**4.2 Stimulating search and retrieval with questions**

We have known for hundreds of years that repetition helps learning. But constant repetition can be distracting and not very stimulating. We have known for hundreds of years that repetition helps learning. But constant repetition can be distracting and not very stimulating. We have known for hundreds of years that repetition helps learning. But constant repetition can be distracting and not very stimulating. We have known for hundreds of years 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; Too much repetition used within the context of learning can be:

1. Good for the soul
2. Bad for the soul
3. Distracting
4. Stimulating

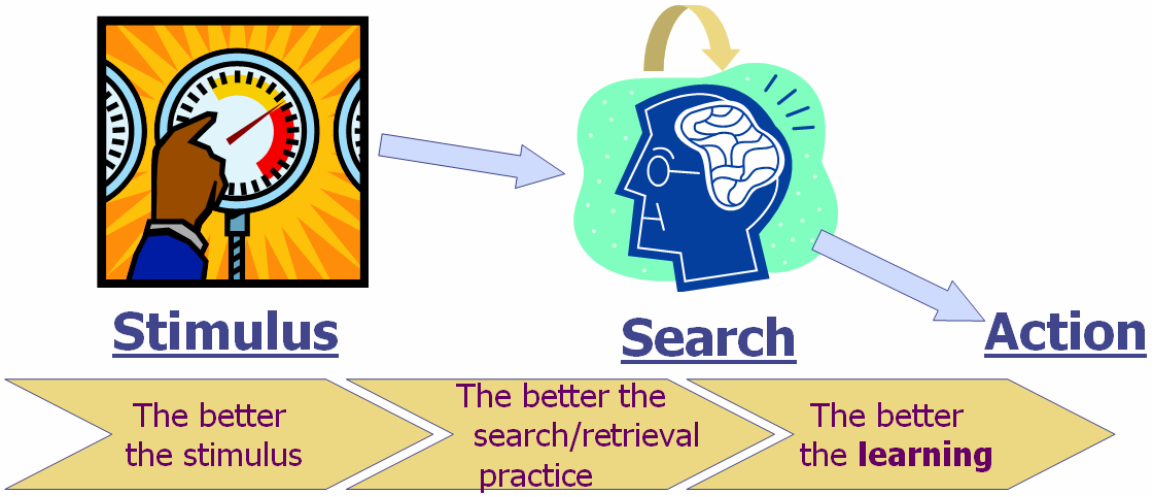
And here is another: Too much repetition used within the context of learning can be:

1. Good for TV advertising
2. Easy to learn
3. Bad for the instructor
4. Not very stimulating

Hopefully you find these examples more stimulating than our first 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.

What works best is to conduct that questioning 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 realistic the stimulus the better because it will simulate the search and retrieval practice that will be used in the performance situation. The environment for learning is a less obvious factor in learning. 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.

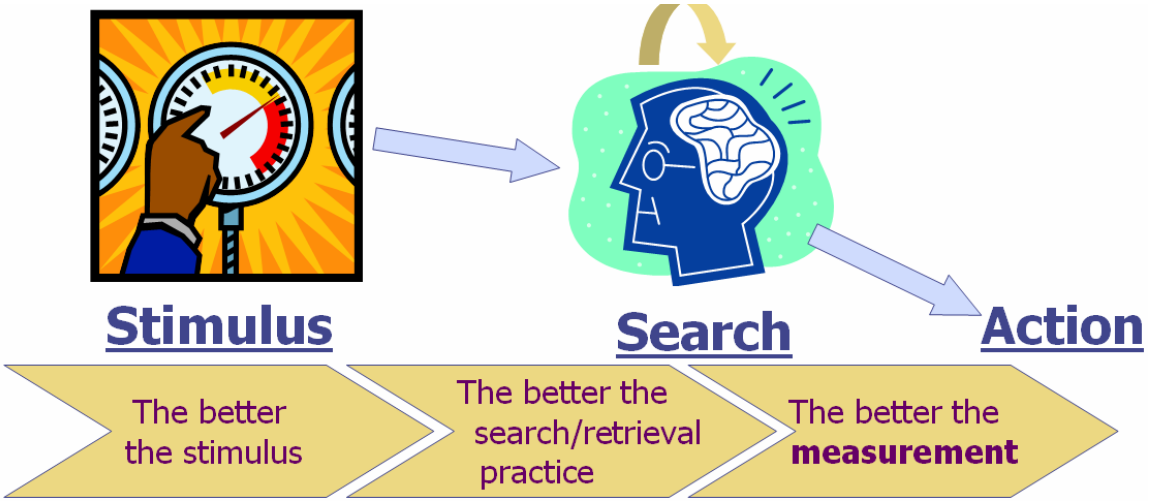

*© Copyright Questionmark Corporation 2004*




The better the stimulus, the better the search and retrieval practice and, therefore, the better the learning. However, sometimes high cost can force a compromise. For instance, if trainers put students in a real-life military aircraft, it's extremely expensive and risky to use hundreds of millions of dollars of equipment to teach someone to fly. But putting that same student in front of a computer with a multiple choice question doesn't really provide an adequate learning experience. However, if the tests use graphics or video on the computer, that greatly improves the stimulus. Here is another, less dramatic 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

Just as we will provide better search and retrieval practice if we position the learning environment to be close to the performance environment, the same rule applies for measurement. The closer the measurement environment is to the performance environment the better the measurement of knowledge and skills will be. That is why a part of the driving test is performed by actually driving a vehicle. This could be expressed as:





© Copyright Questionmark Corporation 2004

## 4.4 Factors that influence what people learn

It's clear that there are a number of factors that influence why people learn in different environments. People's attention wanders. In virtual presentations they might read their e-mail or talk to someone who comes into their office. They are distracted. In a classroom environment, students might be looking out the window or thinking about other subjects.

Secondly, 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 grasp initially. They need to hear the information a couple of times before it starts to sink in.

Even if a leaner absorbs everything there is a good chance they will forget something. Ultimately we can only remember what we can retrieve. The student might forget because the stimulus or environment cues aren't present or because too much time has passed since they learned something. For example, few people use algebraic equations and so have forgotten how to solve them while most high-school students still can.

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

There are many factors that affect how much we learn and how much we can retrieve: the learning environment, the manner in which the material is presented, how that material is taught, as well as others. But 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, training gets boring. Asking questions constitutes another form of repetition that doesn't get boring because it forces students to 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 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 students connect their environment with how to retrieve the required information when they need it.


*© Copyright Questionmark Corporation 2004*

## 4.5 Research illustrates benefit 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 | | |
| Initial | 30% | 110% |
| Subsequent | 15% | 40% |
| Feedback *(to correct misconceptions)* | 15% | 50% |
| Asking questions to provide retrieval practice | 30% | 100% |
| Questions spaced over time (*to reduce forgetting)* | 5% | 40% |
| **Total learning benefit range** | 100% | 380% |

*Based on research by Dr. Will Thalheimer see www.work-learning.com*

For more information or consulting advice on the learning benefits of questions and the benefits of simulation-like questions please refer to Dr. Will Thalheimer at Work Learning Research (www.work-learning.com).


© Copyright Questionmark Corporation 2004

## 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, diagnose 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 are 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 make sense and reports 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 them to the most appropriate learning experience.

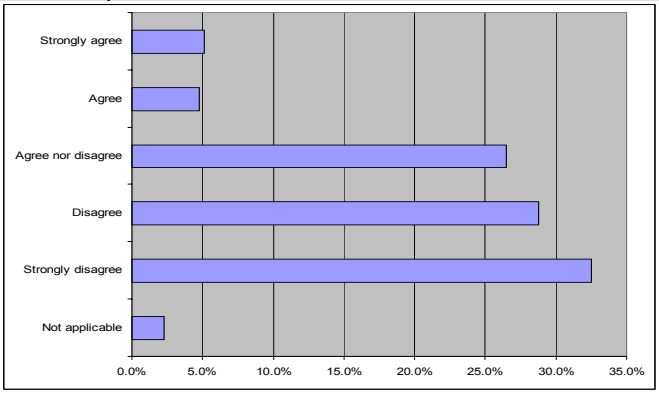
There are two aspects of analyzing a diagnostic assessment:

- 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 students and instructors and determining if students 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 Lykert scale item, "I felt that the course that I was placed in was too advanced for me," clearly show that something is amiss:

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

Times presented: 78
Times answered: 72





© Copyright Questionmark Corporation 2004

## 5.2 Analyzing Formative Assessments

A formative assessment has a primary role of providing search and retrieval practice for a student and as well as prescriptive feedback (item, topic and/or assessment level). In formative assessments students receive feedback at an item level or at a topic level. This helps the person who took the assessment 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 student.

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

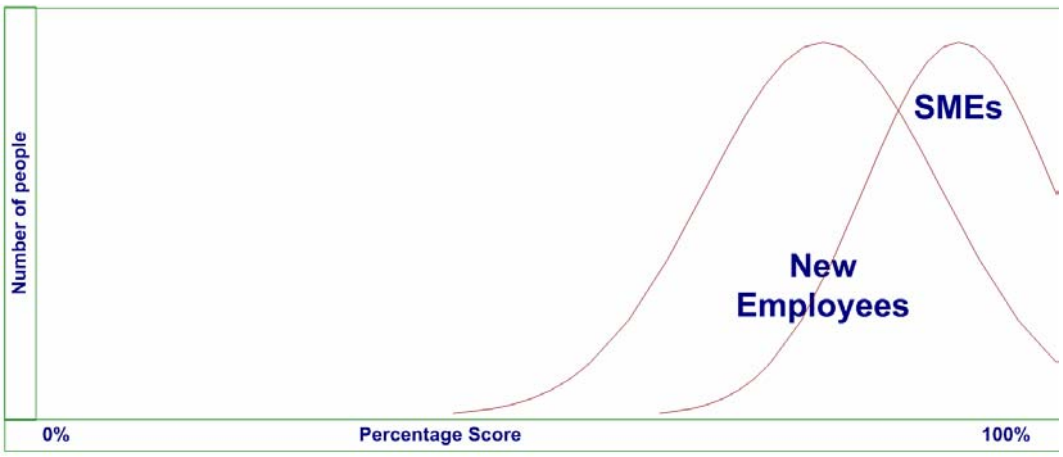
The job task analysis (JTA) survey asks questions of subject matters experts and those on the job to determine the significance and frequency of particular tasks. The JTA guides curriculum design and the development of questions to test the knowledge, skills, abilities, and attitudes that relate to a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

The needs analysis assessment report (below) delivers its scores by topic. When evaluating the groups' scores 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 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 group | Average scores from SMEs | Gap |
|---|---|---|---|
| Food Safety | 60% | 93% | 33% |
| Food Packaging | 90% | 91% | 1% |
| Customer Service | 82% | 71% | 9% |
| Overall Score | 91% | 69% | 22% |

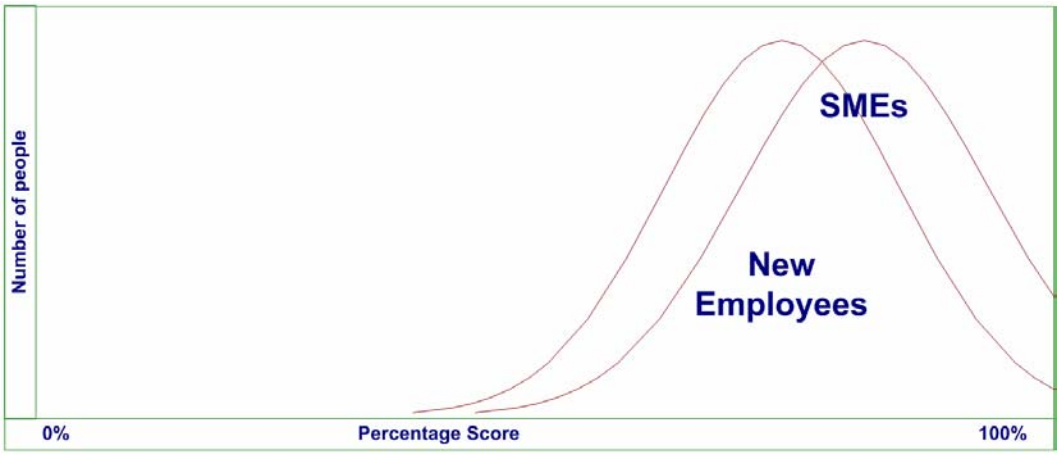
Examining overall results can yield little actionable value. The chart below shows that there is a difference between new employees and SMEs but it is challenging to determine a suitable course of action:


© Copyright Questionmark Corporation 2004




Overall Results

The chart 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.




Results on the topic of Packaging and Presentation

The chart 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 understand that it needs to invest in new hire training to ensure that food is prepared safely.


*© Copyright Questionmark Corporation 2004*


Results on the topic of Food Safety

## 5.4 Analyzing Reaction Assessments

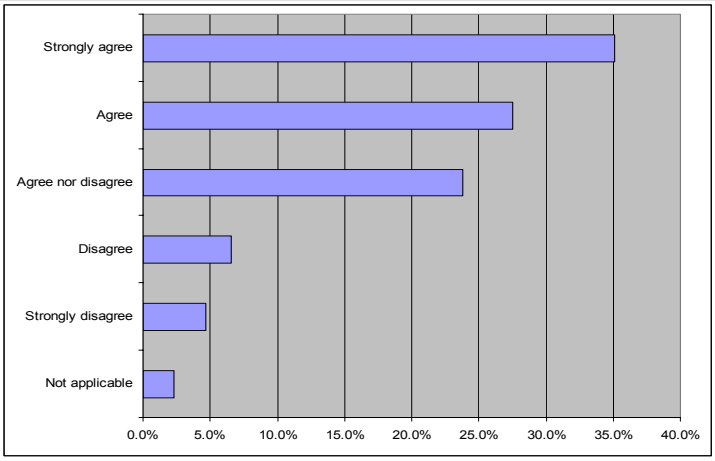
A reaction assessment determines the satisfaction level with a learning event or an assessment experience. These assessments are often known as Level 1 evaluations (as per Dr. Kirkpatrick), course evaluations, smile or happy sheets. They are completed at the end of a learning of certification experience.

Reaction assessments aid the planning process for changing a course.

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.

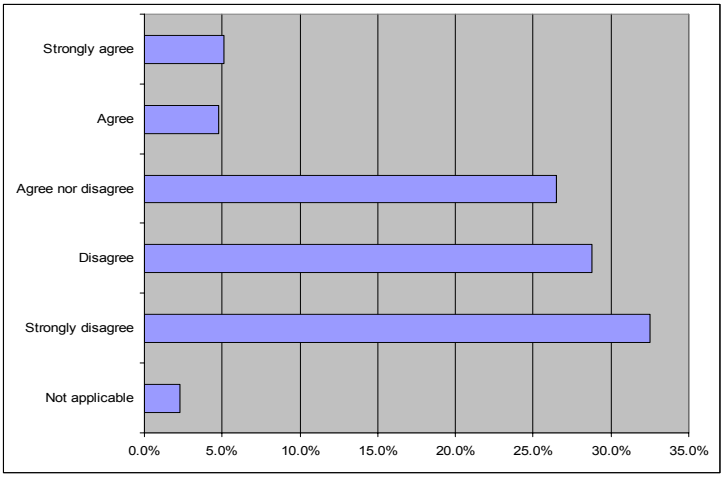
Times presented: 242
Times answered: 227




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


© Copyright Questionmark Corporation 2004

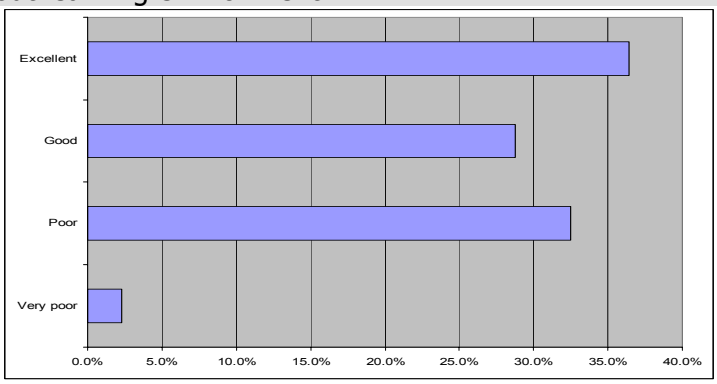
Times presented: 242
Times answered: 227




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.

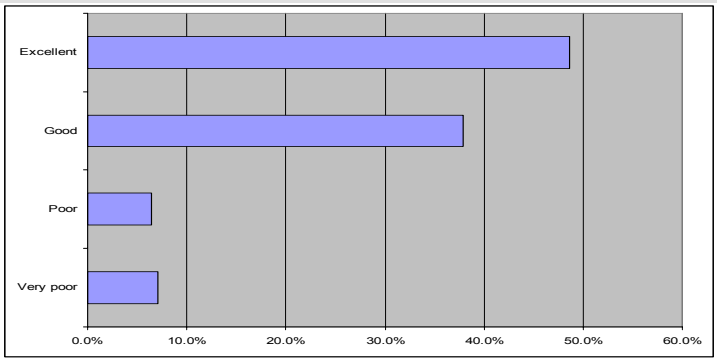
## Aggregation of questions about learning environment

Times presented: 239
Times answered: 208




## Aggregation of questions about instructor

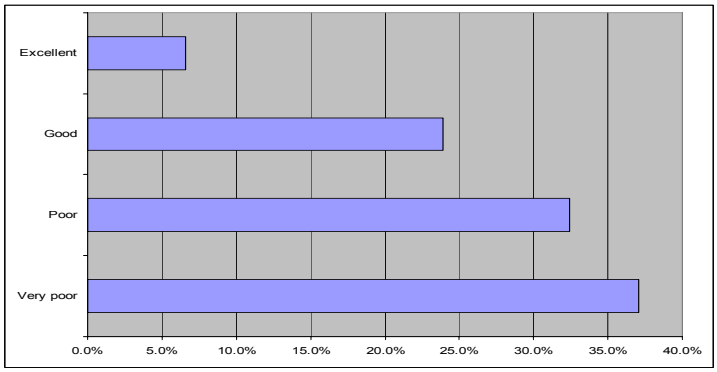
Times presented: 239
Times answered: 208




## Aggregation of questions about the course materials


© Copyright Questionmark Corporation 2004

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 update 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.

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 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, 4 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 that would mean that the question was too easy. It may have produced some learning benefits but probably not. It was just not a very useful question. But if no one got the correct result, that would imply that the question was too difficult and, again, not very useful. Questions that distinguish between knowledgeable candidates and not so knowledgeable candidates are the useful questions.

In a typical multiple choice question, if everybody guessed, approximately 25 percent would select each choice. As a result .25 is normally the lowest difficulty number you'd expect to see there if it's a really poor question with poor choices. Typically, values for difficulty will be in the range of .6 to .8, but these really depend on the cut-score (pass/fail) score of test.


© Copyright Questionmark Corporation 2004

To make a test more difficult it isn't necessary to start thinking up more difficult questions. Begin by ensuring that the questions are valid and provide the right focus and then simply adjust the pass/fail score.

The column titled the "participant mean" averages the final scores on the entire assessment. Of the people who selected choice A, their final scores produced a mean of 36 percent, 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 form -1.0 to + 1.0 and illustrates how the results from this question compared to how students did on the overall test. 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 6$^{th}$ column. In our example the score is +0.49.

Now imagine that a question about baseball is inserted. 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 negative correlation, because this particular question on baseball does not correlate to the overall test results.

The last column in the chart shows a statistic known as Outcome Correlation. This question performed well because the wrong choices have negative correlation (the student didn't do well on the test) and the right choice has positive correlation, which is what assessment designers seek.

### 5.5.2 Setting pass/fail score for norm-referenced tests

Setting the pass/fail score for a norm-referenced test is fairly simple. First, determine how many people should 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 |

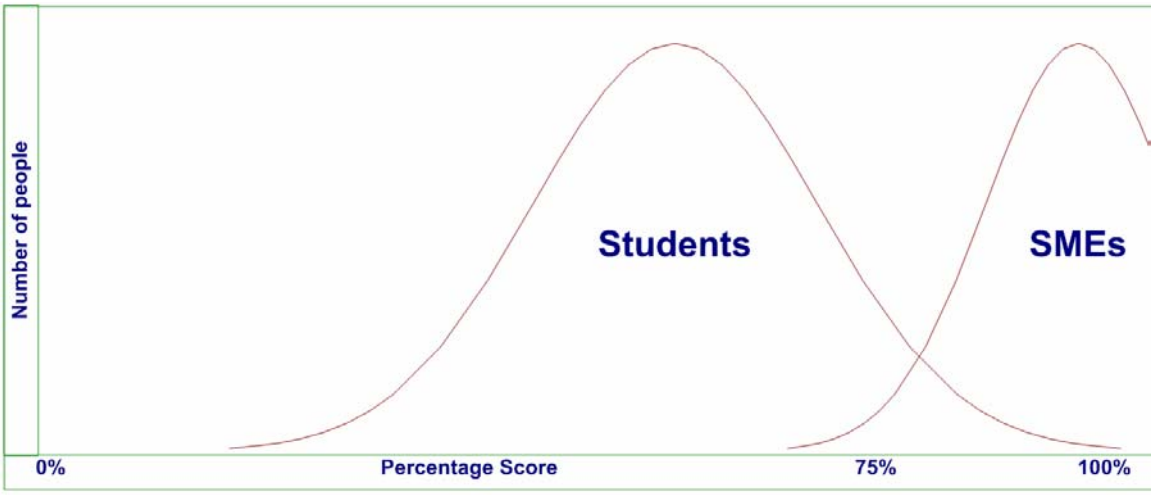

© Copyright Questionmark Corporation 2004

| | |
|---|---|
| 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 students, who are assumed not to know much about a subject, to subject matter experts (SMEs) who do.




These people (SMEs) perform well on the job, and if we look at the overall results a lower number of students performed well compared to the scores that experts achieve. For students the bell curve is weighted toward the lower end of the curve, while for subject matter experts the scores move to the higher end.

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


© Copyright Questionmark Corporation 2004

## 6. The Benefits of Computerizing Assessments

Clearly, assessments of all types can have a major impact upon what and how students 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 very difficult to gather the results needed to effect change in a timely fashion.

By computerizing the assessments it becomes possible to obtain valuable results almost instantly. This rapid feedback improves the quality of learning and satisfaction of the learners. Providing practice exercises with instant feedback greatly enriches the learning process.

Having these kinds of results can have an enormous impact upon the bottom line as well as well as on productivity and regulatory compliance. For example, organizations that must demonstrate regulatory compliance – truth in lending, truth in savings, food safety – can, using computerized assessments, rapidly demonstrate that they've actually complied, thereby satisfying regulators.

To yield real benefits from the assessment techniques described here, 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 students to reach their – and your – objectives.

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

Questionmark can provide the computerized tools to gain effective feedback and effective reports that lead to the results you seek. You can learn more about Questionmark's assessment products from www.questionmark.com.


*© Copyright Questionmark Corporation 2004*

**Recommended reading:**

*Criterion Referenced Test Development: Technical and Legal Guidelines for Corporate Training and Certification* by Sharon A. Schrock and William C. Coscarelli (ISBN 1-890289-09-4)

*Evaluating Training Programs: The Four Levels* by Donald L. Kirkpatrick (ISBN: 1-576750-42-6)

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

*Work Learning Research* white papers by Will Thalheimer (at www.work-learning.com)

White Papers available from the Questionmark web site:

*Delivering Computerized Assessments Safely and Securely*
http://www.questionmark.com/communities/getresource.asp?file=DeliveringComputerisedAssessmentsSecurely.pdf&group_id=5

*The Learning Benefits of Asking Questions* by Dr. Will Thalheimer
http://www.questionmark.com/communities/getresource.asp?file=LearningBenefitsOfQuestions.pdf&group_id=5

*Creating and Deploying Computerized Level 1 Assessments*
http://www.questionmark.com/communities/getresource.asp?file=Level%201%20assessments.pdf&group_id=5

*Improving Training Evaluations in Organizations* by Dr. Paul Squires
http://www.questionmark.com/communities/getresource.asp?file=training_evaluation.pdf&group_id=5


*© Copyright Questionmark Corporation 2004*

## 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 student's knowledge or skills for the purpose of documenting their current level of knowledge or skill. |
| **Test** | A diagnostic assessment to measure a student's knowledge or skills for the purpose of informing the student 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 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 student and to provide prescriptive feedback (item, topic and/or assessment level). |
| **Lykert scale** (Likert) | A method to prompt a respondent to express their opinion on a statement being presented. Lykert scales are often 4 point scales (strongly agree, agree, disagree, strongly disagree), 5 point scales (strongly agree, agree, neutral, disagree, strongly disagree), but sometimes 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 student 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. |
| **SME** | Subject Matter Expert |


*© Copyright Questionmark Corporation 2004*

**About Question*mark*:**

Question*mark* has been producing testing and assessment software since 1988. Businesses, governments, schools, colleges, and universities in more than 50 countries use Question*mark* software. Questionmark has more than 1,800 customers using Perception, with approximately 12,000 authoring systems installed, and systems shipped have been configured millions of participants. Typical applications include exams, quizzes, study aids, course evaluations, surveys, diagnostic tests, pre-course skills assessments, and course evaluations.

**Contact information:**


North America:
Questionmark Corporation
Tel: (800) 863-3950
Fax: (800) 339-3944
Email: info@questionmark.com
Web: http://www.questionmark.com/

Europe:
Questionmark Computing Limited
Tel: + 44 207 263 7575
Fax: + 44 207 263 7555
Email: info@qmark.co.uk
Web: http://www.questionmark.com/




*© Copyright Questionmark Corporation 2004*