question mark

WHITE PAPER

Assessment Results You Can Trust

White Paper

Why trustable assessment results matter.

Learn how Questionmark's assessment management system helps you make your assessments valid and reliable—and therefore trustable.

Authors: John Kleeman
Eric Shepherd

Contents

Executive Summary

Making important decisions about people based on an error-prone process or unreliable data is unwise. This white paper explains the reasons that organizations target assessment results that they can trust.

Assessment results must be both valid (measure what you are looking for them to measure) and reliable (consistent in that measurement). This white paper explains the features required in an assessment management system to assist organizations in developing, delivering and reporting on assessments to ensure that they are valid and reliable.

This white paper draws upon the metaphor of a doctor using results from a blood test to diagnose an illness and then prescribing a remedy. If the wrong test is ordered by the doctor, delays will occur, or if the results are untrustworthy, serious consequences could result. Using this metaphor, it is easy to understand the personnel and organizational risks that can result from making decisions based on untrustworthy results. If an organization assesses someone’s knowledge, skill or competence for health and safety or regulatory compliance purposes, you need to ensure the assessment instrument is designed correctly and runs consistently.

Engaging subject matter experts to generate questions to measure the knowledge, skills and abilities required to perform the essential tasks of the job is crucial in creating the initial pool of questions. However, subject matter experts are not necessarily experts in writing good questions, so having a quality control process that allows assessment experts (e.g., instructional designers or psychometricians) to easily review and amend assessment items is a key feature required of any authoring system.

For assessments to be valid and reliable, it is necessary that you follow structured processes at each step from planning through authoring to delivery and reporting. In this white paper, we explain why it is important to follow appropriate processes when planning the assessment, authoring items, assembling the assessment, trialing it, and then when delivering and reporting on it.

Using an assessment management system to organize your assessments makes it easier to create valid and reliable, and therefore trustable, assessments instead of using simpler tools. Our advice is simple—if you need to use assessment results to make business decisions or decisions about people, use a system that you can trust.

We hope this white paper will be useful to you if you are using Questionmark or if you are considering purchasing Questionmark and need to understand its value compared to other software and systems that you have.

1. Introduction

This white paper has been written to help corporate and government stakeholders create, deliver and report on assessments to produce trustable results that can effectively measure the competence of employees and their extended workforce.

In any organization, an employee in a job role performs multiple tasks. Each task requires knowledge, skills and abilities in order to perform that task successfully. Almost every organization needs to assess that their workforce has these capabilities both for their own business purposes and to satisfy the needs of regulators.

Most organizations use assessments for recruitment, onboarding, promotion or other talent development initiatives. They also need to assess their employees as part of regulatory compliance and to ensure the health and safety of the workforce. The stakes of these assessments are substantial: decisions made based on the results can impact both the reputation and financial well-being of the organization, as well as life, limb and livelihood.

Assessments are especially needed when employees learn informally or on the job. There is a growing understanding that most workplace learning happens in the 70:20:10 model, where 70% of learning is by doing, 20% of learning is from peers, and 10% of learning is from formal study. Nonetheless, regardless of how people learn, they will want to know “they got it,” and so will you. The best and often the only way to ensure that someone in a job role has the knowledge, skills and abilities to do the tasks in it is via an assessment.

Whatever you use assessments for, this white paper explains why it is important to be able to trust assessment results if you are making decisions about people. As we will explain, the key to trustable results is reliability (predictable assessments that return the same results if taken more than once) and validity (that they measure what you are seeking to measure).

There are other kinds of assessments, but in this white paper, we focus on assessments that measure knowledge, skills and abilities within the workplace, sometimes called “competency tests.” Such tests are typically criterion referenced—they measure someone’s competence against agreed-upon criteria and usually give a pass or fail result. Trustable results for these assessments can only come from reliable and trustable processes.

Most learning management systems (LMSs) and e-Learning creation tools have the capability to create basic quizzes and tests, and it can be tempting to use these to create assessments with which to make decisions about people. However, this white paper explains why such systems usually fail to automate, manage, and track key processes related to ensuring valid, reliable and trustable assessment results. Assessments constructed in such tools can be useful for learning, but if you want trustable assessments for other purposes, you should consider an assessment management system like Questionmark's.

For example, many LMSs and e-Learning tools require specialists or experts to use them to create content. However, the best way to get great assessment items is to harvest or "crowdsource" them from subject matter experts and involve these experts in the review process. Many such tools put assessments and their questions alongside learning content, but in order to manage assessments effectively, you need an item bank to hold, index, review and update questions.

This white paper explains how an assessment management system like Questionmark's will help you obtain trustable assessment results.

2. Why trustable assessment results matter

It takes effort to write questions, and it costs time for your employees to answer them—this cost is only worth it if you can trust the results. An assessment score or pass/fail result without good practice behind it is both meaningless and dangerous.

If assessment results are trustable, you can:

- Make better decisions in talent management—recruiting, promoting and developing talent;
- Reduce regulatory compliance risk, errors and fines by identifying poor practices and lack of knowledge before it impacts your business or causes regulatory fines;
- Reduce risks to personnel by ensuring that people are competent in health and safety procedures;
- Provide evidence, in a court of law, that demonstrates that your organization assessed employees' competencies to ensure compliance with regulations and health and safety procedures;
- Encourage ownership of the key competencies required for a job;
- Ensure your workforce is competent and gain the business impact from this in customer service, manufacturing quality and any other aspects of your business that rely on people;
- Identify training needs to use resources wisely to train effectively and not waste employee time;
- Onboard new employees well;
- Set up partner programs to verify knowledge and skills of your sales and technical channels to verify and enhance partner skills;
- Develop certification programs for your employees, customers or partners that add value to your product offering; and
- Use "big data" technology to correlate assessment results with performance, which is only possible if you can trust your assessment results.

As a corollary, if your assessment results are not trustable, then you run the risk that you will:

- Hire the wrong people;
- Promote the wrong people;
- Run a legal risk from making employment decisions that are not defensible;
- See errors in health and safety, manufacturing, customer service and regulatory compliance that lead to reputation loss and/or compliance fines;

- Have little or no evidence that employees were competent to perform tasks safely and in compliance with regulations;
- Waste time in training people what they already know;
- Fail to train people in what they need to know;
- Devalue any partner or certification programs that use assessments;
- Draw incorrect conclusions from assessment results—i.e., "garbage in means garbage out"; and
- Waste the time spent in creating assessments and delivering them.

Unreliable, invalid assessments can help a bit in learning. Giving people questions as retrieval practice helps prevent forgetting. And if people know that there is a test or quiz after learning, it can encourage them to study for the test or quiz even if the test is poorly designed. However, for any kind of decision-making based on assessments, you need to be able to trust the results, which, as we will explain in the next section, means that you need **reliable** and **valid** assessments.

3. Validity and reliability are the keys to trust

How can you trust assessment results? The two keys are reliability and validity.

Reliability explained

An assessment is **reliable** if it measures the same thing consistently and reproducibly. If you were to deliver an assessment with high reliability to the same participant on two occasions, you would be very likely to reach the same conclusions about the participant's knowledge or skills. A test with poor reliability might result in very different scores across the two instances.

20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

Not reliable

An unreliable assessment does not measure anything consistently and cannot be used for any trustable measure of competency. It is useful visually to think of a dartboard; in the diagram to the right, darts have landed all over the board—they are not reliably in any one place.

In order for an assessment to be reliable, there needs to be a predictable authoring process, effective beta testing of items, trustworthy delivery to all the devices used to give the assessment, good-quality post-assessment reporting and effective analytics.

We explain key ways to ensure reliability later in this white paper.

Validity explained

Being reliable is not good enough on its own. The darts in the dartboard in the figure to the right are all consistently in the same place, but not in the right place. A test can be reliable but not measure what it is meant to measure. For example, you could have a reliable assessment that tested for skill in word processing, but this would not be valid if used to test machine operators, as writing is not one of the key tasks in their jobs.

Reliable but not valid

An assessment is **valid** if it measures what it is supposed to measure. So if you are measuring competence in a job role, a valid assessment must align with the knowledge, skills and abilities required to perform the tasks expected of a job role (this is called “content validity”). In order to show that an assessment is valid, there must be some formal analysis of the tasks in a job role and the assessment must be structured to match those tasks. A common method of performing such analysis is a job task analysis, which surveys subject matter experts or people in the job role to identify the importance of different tasks.

As well as having content validity, it is also important that an assessment has “face validity.” Face validity is the perception amongst non-experts that the test measures what it is intended to measure. Stakeholders in the assessment process (consumers, managers, employees, participants, regulators) will only trust the results of the assessment if it appears to be appropriate.

Assessments must be reliable AND valid

Trustable assessments must be reliable AND valid. The darts in the figure to the right are in the same place and at the right place. When you are constructing an assessment for competence, you are looking for it to consistently measure the competence required for the job.

Reliable and valid

The concepts of reliability and validity are cornerstones of the assessment world and are well enshrined in international ISO standards. For example, ISO 17024[1] states: “Certification of a person should be based on objective evidence obtained by the certification body through a fair, valid and reliable assessment.”

[1] ISO/IEC17024:2012 Conformity assessment – General requirements for bodies operating certification of persons

A comparison with blood tests

It is helpful to consider what happens if you go to the doctor with an illness. The doctor goes through a process of discovery, analysis, diagnosis and prescription. As part of the discovery process, sometimes the doctor will order a blood test to identify if a particular condition is present, which can diagnose the illness or rule out a diagnosis.

It takes time and resources to do a blood test, but it can be an invaluable piece of information. Doctors could order every test; however, in practice they isolate the most probable causes of an ailment and then order tests to identify a specific issue. For example, one type of blood test might measure the glucose level in blood—a high level of glucose can be a sign of diabetes. Another type of blood test measures the amount of troponin in blood—a high level of troponin can mean that a heart attack is imminent or has happened.

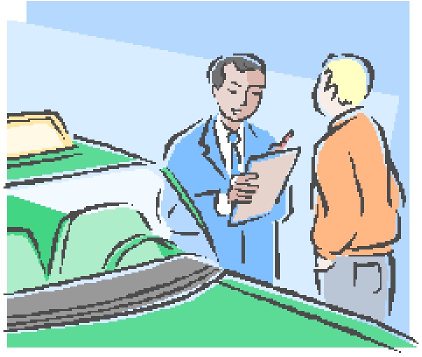
Within the world of health care, a great deal of effort goes into making sure that blood tests are both reliable (consistent) and valid (measure what they are supposed to measure). For example, just like exam results, blood samples are labelled carefully, as shown in the picture[2], to ensure that patient identification is retained.

A blood test that was not reliable would be dangerous—a doctor might think that a disease is not present when it is. Furthermore, a reliable blood test used for the wrong purpose is not useful—for example, there is no point in having a test for blood glucose level if the doctor is trying to see if a heart attack is imminent.

The blood test results are a single piece of information that helps the doctor make the diagnosis in conjunction with other data from the doctor's discovery process. In exactly the same way, a test of competence is an important piece of information to determine if someone is competent in their job role. However, that test is only useful if it is reliable and valid.

[2] Picture by Graham Colm

A comparison with driving tests



Would road safety increase or decrease if we abolished the driving test? Most likely, it would become less safe on the roads, as the driving test is typically valid and reliable. It is valid as it requires you to drive a real car under supervision, usually with specific practical maneuvers required. In addition, effort is put into making it reliable, for instance by setting a standardized procedure all candidates must go through and ensuring that the examiners are trained to score tests in the same way. Therefore, you should pass or fail the driving test based on how well you did objectively, not on the subjective judgment of the examiner.

A concern in some countries in recent years has been that the driving test does not deal well enough with hazards and emergencies, as it is not safe to present real hazards (like a person running out into the road in front of a car) during a test. Dealing with emergencies is an important competence for a driver, but not one that is safe to measure when actually driving. There has been a concern that the test is less valid because of this. Therefore, some countries have introduced an additional element of the test using videos to test hazard perception. This likely makes the driving test more valid.

Why reliability and validity matter for workplace assessments

Modern organizations need to rely on their people being competent.

Would you be comfortable in a high-rise building that was designed by an unqualified architect?

Would you fly in a plane if the pilot had not passed a flying test?

Would you let someone operate a machine in your factory if they did not know what to do if something went wrong?

Would you send a salesperson out to make a sale if they did not know what your products do?

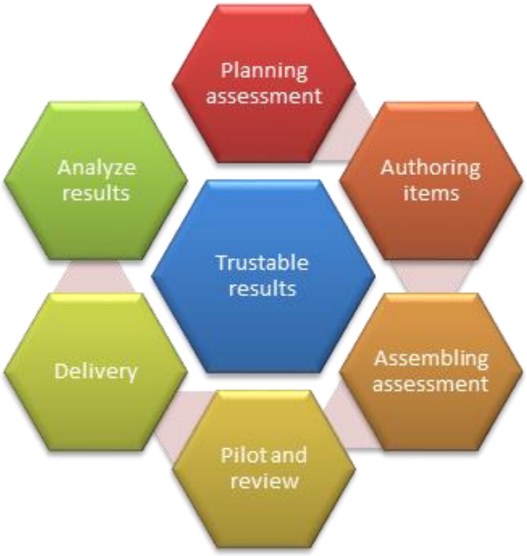
Can you demonstrate to a regulatory authority that your staff is competent and fit for assigned jobs if you do not have trustable assessments?

In all these cases and many more, it is essential that you have a reliable and valid test of competence. If you do not ensure that your workforce is qualified and competent, then you should not be surprised if your employees cause accidents, get your organization fined for regulatory infractions, give poor customer service or cannot repair systems effectively.

4. How Questionmark solutions help deliver valid and reliable assessments

So far in this white paper, we have explained why it is important for assessments used in the workplace to be valid and reliable if they are to be trustable. In this next part of this paper, we will share why using Questionmark technology makes it more practical and realistic to create and deliver valid and reliable assessments than if you use a more basic tool.

The key to validity and reliability starts with the authoring process. If you do not have a repeatable, defensible process for authoring questions and assessments, then however good the other parts of your process are, you will not have valid and reliable assessments. It is useful to think of six aspects of the assessment process, as shown in the diagram below.



You need to start with planning the test or blueprinting, which is working out what it is that the test covers. Then you author the items and assemble them for use in a test. After piloting and reviewing to check that the test is valid and reliable, you move on to delivering it to participants and analyzing the results. Each step contributes to the next, and useful analysis of the results is only possible if **every** previous stage has been done effectively.

You also need security throughout all processes, as a failure of security can also risk the trustworthiness of the results. The following sections describe some key capabilities in each area that you should be looking for in an assessment management system.

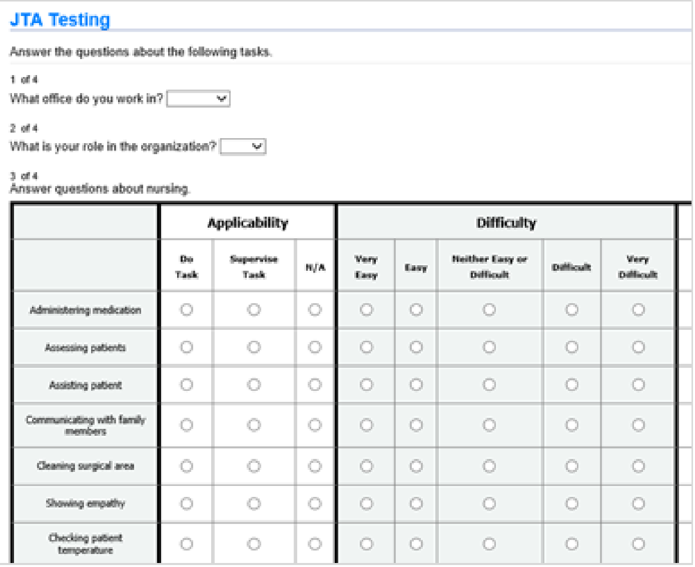
Planning the assessment

The critical value that Questionmark brings is its structured authoring processes, which enable effective planning, authoring, and reviewing of questions and assessments and makes them more likely to be valid. It is obvious, but if you do not plan the assessment properly to match the tasks that people require in a job, then however carefully you construct and deliver the assessment, it will not measure the competence properly.

Some of the most important capabilities of Questionmark in planning the assessment are:

1. **Use job task analysis surveys to help blueprint assessments**
   In order for an assessment to be valid, it must match the job role or competency that it is testing.

   Job task analysis (JTA) surveys are used to analyze what tasks within a job role are most important and are a key way for you to check what topics need to be covered in an assessment. Typically you survey "masters"—people doing a job already or who are already certified in tasks in their job that are the most important and done most frequently—to determine the areas of questioning. They provide evidence that the coverage of questions matches the coverage of what is needed to do a job.



JTA Testing

Answer the questions about the following tasks.

1 of 4
What office do you work in?

2 of 4
What is your role in the organization?

3 of 4
Answer questions about nursing.

	Applicability			Difficulty				
	Do Task	Supervise Task	N/A	Very Easy	Easy	Neither Easy or Difficult	Difficult	Very Difficult
Administering medication	○	○	○	○	○	○	○	○
Assessing patients	○	○	○	○	○	○	○	○
Assisting patient	○	○	○	○	○	○	○	○
Communicating with family members	○	○	○	○	○	○	○	○
Cleaning surgical area	○	○	○	○	○	○	○	○
Showing empathy	○	○	○	○	○	○	○	○
Checking patient temperature	○	○	○	○	○	○	○	○

   Often a JTA will ask respondents about the applicability, difficulty and/or frequency of each task. Questionmark technology offers a JTA question type and provides JTA reports to help you run JTAs easily and effectively and get useful data to use in your assessment design[3].

[3] For more on JTAs, see the Questionmark blog at https://www.questionmark.com/create-a-reliable-test-with-jta

If you do not use JTAs, you risk that the assessment will not cover the right competencies and tasks, and so it will be invalid and also indefensible if challenged.

2. **Organize items in an item bank with topic structure**
   Once the JTA has been completed, you can determine the topics that an assessment needs to cover. There are huge benefits to using an assessment management system with an item bank that structures items by hierarchical topics as this facilitates:
   - An easy management view of all items and assessments under development
   - Mapping of topics to relevant organizational areas of importance
   - Clear references from items to topics
   - Use of the same item in multiple assessments
   - Simple addition of new items within a topic
   - Easy retiring of items when they are no longer needed
   - Search capabilities—for example, identifying questions that need updating when laws change or a product is retired

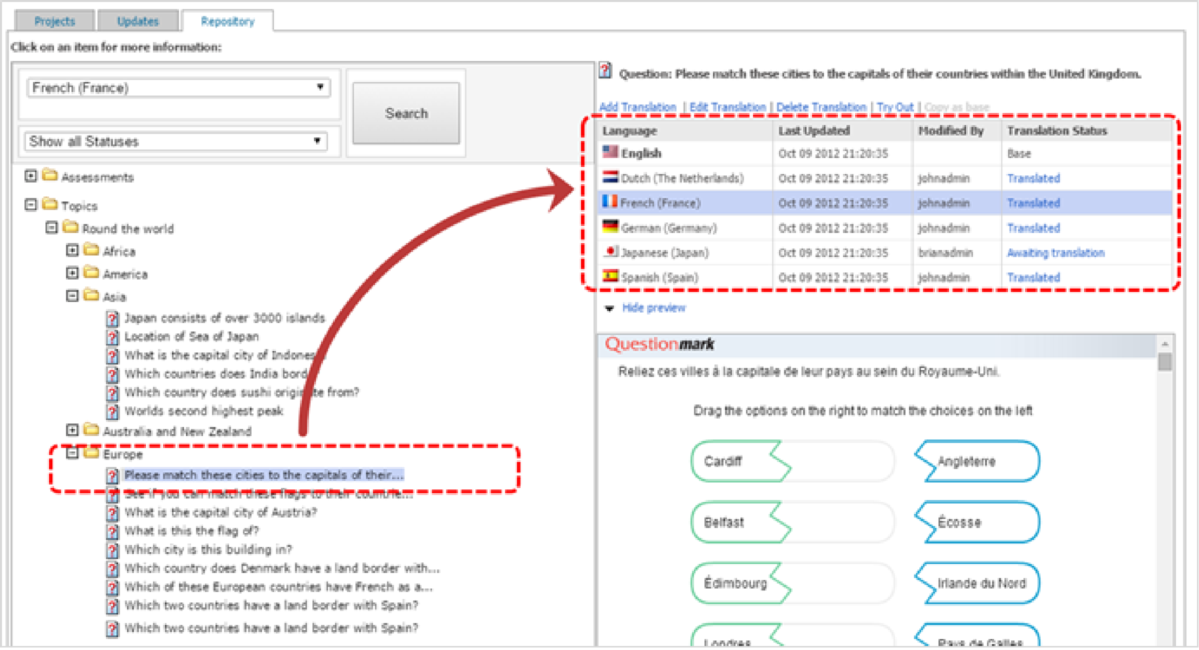
   Some standalone e-Learning creation tools and some LMSs do not provide you with an item bank, but require you to insert questions individually within an assessment. If you only have a handful of assessments or you rarely need to update assessments, such systems can work, but for anyone with more than a few assessments, you need an item bank.

3. **Metatag items to associate with task knowledge, skill, or ability**
   As well as organizing items by topics, it can be helpful to use metatags to index items in other classifications, including specific job tasks, knowledge, skills and abilities. This allows more effective management of items and selection within the appropriate assessments.

4. **Authoring processes secure to protect against content theft**
   If item or assessment content leak out during the assessment construction process, then validity will be lowered, as people can see the questions in advance. You need individual logins for authors that are well protected by strong passwords and good policies and culture within your team to prevent this. You also need a way to easily set up differential security at the topic or sub-topic level so that authors and reviewers only see the questions they need to and not the entire item bank.

5. **Manage translations to aid multilingual assessments**
   If you have a need to deliver the same assessment to people who speak different languages, you need translation and multilingual delivery capabilities. Obviously, if someone does not understand the language that the assessment is in, it will not validly test their competence. Questionmark translation management aids you in translating items and assessments into multiple languages.

   For example, the screenshot below shows a sample question translated into five languages.



Click on an item for more information:

- Assessments
- Topics
  - Round the world
    - Africa
    - America
    - Asia
      - Japan consists of over 3000 islands
      - Location of Sea of Japan
      - What is the capital city of Indones
      - Which countries does India bord
      - Which country does sushi origi te from?
      - Worlds second highest peak
    - Australia and New Zealand
    - Europe
      - Please match these cities to the capitals of their...
      - See if you can match these flags to their countrie...
      - What is the capital city of Austria?
      - What is this the flag of?
      - Which city is this building in?
      - Which country does Denmark have a land border with...
      - Which of these European countries have French as a...
      - Which two countries have a land border with Spain?
      - Which two countries have a land border with Spain?

Question: Please match these cities to the capitals of their countries within the United Kingdom.

Language	Last Updated	Modified By	Translation Status
English	Oct 09 2012 21:20:35		Base
Dutch (The Netherlands)	Oct 09 2012 21:20:35	johnadmin	Translated
French (France)	Oct 09 2012 21:20:35	johnadmin	Translated
German (Germany)	Oct 09 2012 21:20:35	johnadmin	Translated
Japanese (Japan)	Oct 09 2012 21:20:35	brianadmin	Awaiting translation
Spanish (Spain)	Oct 09 2012 21:20:35	johnadmin	Translated

Reliez ces villes à la capitale de leur pays au sein du Royaume-Uni.

Drag the options on the right to match the choices on the left

Cardiff	Angleterre
Belfast	Écosse
Édimbourg	Irlande du Nord
Londres	Pays de Galles

Authoring items

Here are some of the key technology capabilities you should be looking for when authoring items.

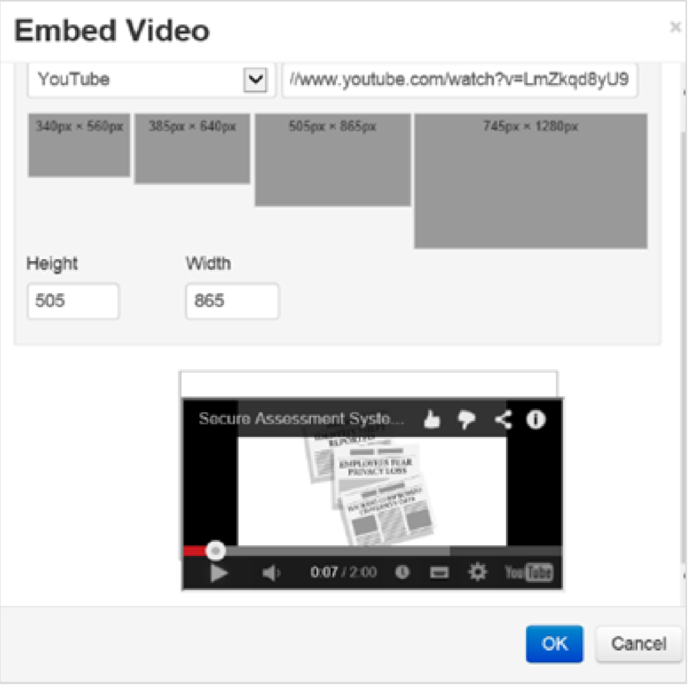
1. **Authoring tool subject matter experts can use directly**
   One of the critical factors in making successful items is to get effective input from subject matter experts (SMEs), as they are usually more knowledgeable and better able to construct and review questions than learning technology specialists or general trainers. If you can harvest or "crowdsource" items from SMEs and have learning or assessment specialists review them, your items will be of better quality.

In addition, you need to review items by SMEs before using them to check that they are accurate.

The Questionmark Live tool from Questionmark is an easy-to-use, browser-based system that allows SMEs to directly login and create or review items. Questionmark Live is widely used; at the time of writing, SMEs and other authors typically create 40,000 questions each month within Questionmark Live.

2. **20+ question types**
   It is better to ask questions that check how people can apply knowledge in the job context rather than just whether they have specific knowledge. Less IT-literate workers may find that “hotspot” and “drag and drop” question types make it easier for them to respond in the context of them performing their tasks without having to develop “office” style typing skills.

3. **Ability to use relevant stimulus including video, audio and equations**
   Effective question stimulus, such as inserting audio and video clips or mathematic formulas, can help simulate conditions encountered on the job that will yield more accurate measurements of someone’s ability to perform. Such stimulus can be useful to provide relevant business or organizational context as well as real-life situations that might be encountered on the job. For example, a video of someone promoting a dishonest act is easier to understand than a textual description of the same.



Embed Video

YouTube

//www.youtube.com/watch?v=LmZkqd8yU9

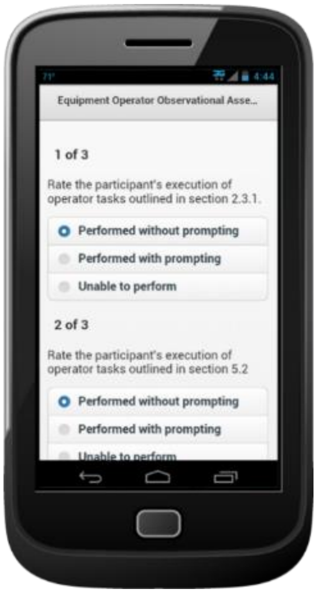
Height
505

Width
865

It is easy to include these in Questionmark within question stimulus, choices, or feedback. The screenshot to the right shows the embedding of a YouTube video within a question.

4. **Use observational items/assessments to help measure performance**

   When assessing practical and communication skills, it is likely more valid to have someone observe the performance of that skill, in conjunction with or instead of asking cognitive questions. For example, if you need to assess someone's ability to weld a joint, perform a medical procedure or interview a customer, having an observational assessment allows you to directly assess the practical skill.



Equipment Operator Observational Asse...

1 of 3

Rate the participant's execution of operator tasks outlined in section 2.3.1.

● Performed without prompting
○ Performed with prompting
○ Unable to perform

2 of 3

Rate the participant's execution of operator tasks outlined in section 5.2

● Performed without prompting
○ Performed with prompting
○ Unable to perform

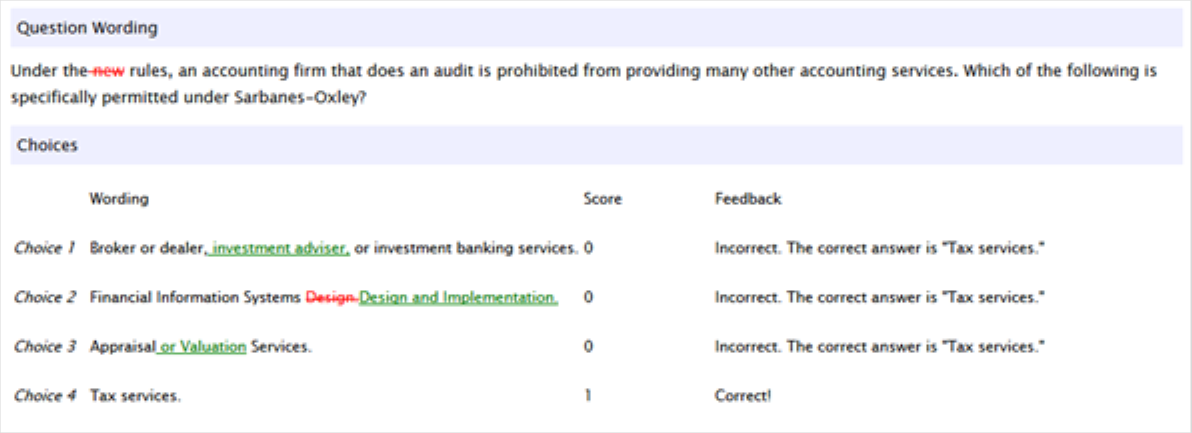
   Questionmark allows you to create observational items, which are often checklists or objective scales of performance. You then put these into an observational assessment, where a supervisor or instructor logs in as an observer and observes and rates the participant on his/her performance. It is common to use tablets or smartphones for such observational assessments, as they can easily be taken into the workplace.

5. **Item version history for legal defensibility**

   It is important that if you are challenged by a regulator, by an employee who feels that the test is unfair, or even in a court of law, you are able to defend that your assessment processes are fair. One key part of this is to show the history of item development to demonstrate the review that happened to make each question. Questionmark records each version of an item and also comments as items are changed. If you ever need to show a regulator or prove in court the process and steps of how an item was constructed and validated, evidence is available to show.

6. **Easy collaboration for item reviewers to help make items more valid**

   Item version history is also used to make it effective for SMEs and other stakeholders to review items. Every time an item is changed, a comment can be stored, and when reviewing item history, it is easy to see comments and past changes and to make a previous version current. There is also a "track changes" capability that allows you to easily see the changes made between versions.



Question Wording

Under the ~~new~~ rules, an accounting firm that does an audit is prohibited from providing many other accounting services. Which of the following is specifically permitted under Sarbanes-Oxley?

Choices

	Wording	Score	Feedback
Choice 1	Broker or dealer, investment adviser, or investment banking services.	0	Incorrect. The correct answer is "Tax services."
Choice 2	Financial Information Systems ~~Design.~~ Design and Implementation.	0	Incorrect. The correct answer is "Tax services."
Choice 3	Appraisal or Valuation Services.	0	Incorrect. The correct answer is "Tax services."
Choice 4	Tax services.	1	Correct!

7. **Search questions to identify those that need updating**
   You need an easy way to search the item bank to find particular questions, especially if there is a change in circumstances that could impact several topics.

8. **Retire questions that are no longer useful**
   When a product or a law changes or a question becomes out-of-date for other reasons, it is useful to "retire" a question, rather than deleting it. Retiring a question means that it stays within the item bank so that related history and reports are available, but it is no longer used in any current assessments.

Assembling the assessment

Once items are created, here are the key capabilities of Questionmark in assembling the assessment for validity and reliability:

1. **Rules-based or random selection of questions from item bank**
   A common approach to constructing competency tests is to use rules to select items at random from topics in an item bank, with the selection and weighting of topics guided by a blueprint. Such a blueprint will often be derived from JTA surveys. Providing you balance the difficulty of items within a topic, this makes an assessment that is different each time it is delivered. This makes it harder for someone to cheat by getting questions or answers from another, as each person's assessment is different. It also allows easy, ongoing management of the assessment—new questions can be added and older questions retired without impacting ongoing delivery.

2. **Random ordering of questions and choices**
   Whether or not you randomize question selection, it is useful to shuffle the order of questions and the order of choices in multiple choice and similar questions—again, to make it harder for validity to be reduced by people communicating with each other about the assessment.

3. **Set a pass score including topic pre-requisites**
   We provide some links to good practice at the end of this white paper, which include advice on setting a pass score fairly. One very useful capability within Questionmark is the ability to make a pass score dependent on topic pre-requisites.

   For example, if an assessment has three topics, you can require that a pass only applies if the participant scores a good enough score in each topic as well as on the assessment as a whole. This ensures that people do not pass tests when they are strong in some topics but weak in a crucial one and helps make an assessment more valid for competence in a job role.

4. **Topic scoring and feedback**
   We recommend that you share topic scores with participants and managers, as if topics are well chosen, they are valuable input in identifying weaknesses and areas for remediation. Topic feedback also helps signpost participants to resources to improve weaknesses. Topic scoring and feedback help face validity, as they show that the pass or fail result is based on solid, topic-by-topic evidence.

5. **Set assessment time limit (and override if required)**
   To make an assessment reliable, it is usual to set a time limit. Questionmark software lets you do this, and it also allows you to vary the time limit when scheduling the assessment to accommodate disabilities. Questionmark also has capabilities to manage or reset the time limit when assessments are disrupted by technical problems.

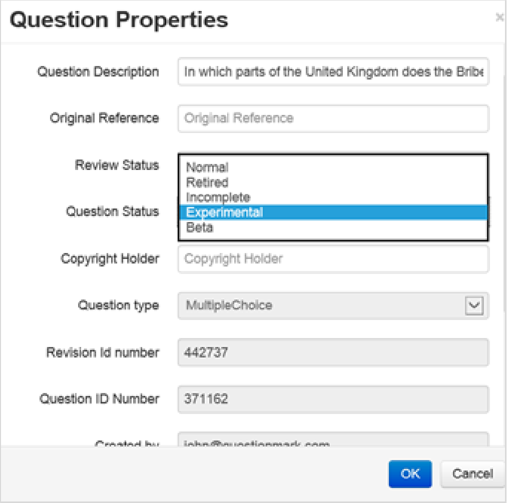
Pilot and review

It is critical to pilot or field test assessments and review the results of the pilot prior to using an assessment. There are a variety of methods of piloting, including “beta” programs, getting instructors or other experts to take the assessments, and including experimental questions within production assessments (not counting towards the score). Key capabilities within Questionmark for pilot and review are:

1. **Easily deliver assessments for trial purposes**
   It requires just a few clicks within the Questionmark user interface to release an assessment for trial purposes. You can easily try out a complete or partially complete assessment in low-stakes environments over the web. Such assessments can be tried out either directly within the Questionmark system or via an LMS or other gateway system. Trial assessments can easily be changed for the trial context, e.g., with a longer time limit, different instructions to participants or just a subset of questions.

2. **Participants can comment on questions and provide feedback to improve validity**
   During review phases, it is useful to turn on comments within questions, allowing participants to add text comments; for example, if they feel a question is ambiguous or unfair. Such comments can easily be viewed within Questionmark reports and provide a simple way for participants to offer comments in context, which helps make the assessments more valid. It is then straightforward to turn off the capability to give comments when moving to a production exam.

3. **Use experimental questions in live assessments without impacting scores**
   Within the Questionmark item bank, you define a status for each question—one of Normal, Retired, Incomplete, Experimental or Beta, as shown in the screenshot on the right.



Question Properties

Question Description	In which parts of the United Kingdom does the Bribe
Original Reference	Original Reference
Review Status	
Question Status	Normal Retired Incomplete Experimental Beta
Copyright Holder	Copyright Holder
Question type	MultipleChoice
Revision Id number	442737
Question ID Number	371162
Created by	john@questionmark.com

   Experimental questions appear normally to a participant but do not count toward the total score or pass mark of an assessment. This allows you to add experimental questions into a production assessment and collect item statistics on them, but not have them impact the reliability and validity of the assessment until those statistics show that they are appropriate.

It will improve the reliability and validity of your assessments if you can try out questions with a real audience before relying on them, and this ability to use experimental questions is an effective way of getting that trial.

4. **Run useful reports on the results of trial assessments**
   All the Questionmark reports, including those for item and test analysis, can easily be run on trial sets of results as well as production results and are effective in reviewing the results of the trial. See the "Analyze results" section below for more on these reports.

Delivery

Once an assessment has been planned, created and piloted, you can deliver it to your participants. It has been shown to be valid and reliable in trial, and you need to ensure that it remains valid and reliable when being delivered for real. Here are some key capabilities of Questionmark that help deliver valid and reliable assessments.

1. **Consumer-quality participant user interface**
   Employees expect business software to have the same quality of user interface as consumer software. For your assessments to have face validity, they need to have a high-quality user experience. It is also important that the questions are clear and easy to navigate through so that test error does not creep in due to confusion or ambiguity on how to answer them. If test delivery is of poor quality, this will discredit the test and reduce face validity.

   Assessments must use responsive design, which means that they adapt to the device/screen being used. Responsive design makes it easy to deliver assessments to different types of smartphones, tablets and other mobile, touch and multi-touch devices. Test takers are often anxious, and it is important that the technology simply works rather than having to make excuses for incompatible browsers and/or devices. Responsive design also accommodates screen size, as you do not want huge navigation controls on a small screen or tiny controls on a large screen. By using responsive design, you can author an assessment once, schedule it once and then deliver it in as many different ways as you like.

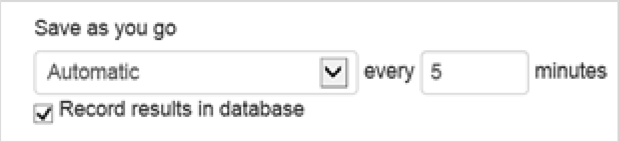
   When delivering an assessment, Questionmark auto-senses the participant's device and browser and then delivers the assessment formatted appropriately for the device's and browser's requirements using responsive design. Questionmark's "on the fly" auto-sizing dynamically adjusts and sizes the assessment's navigation buttons, controls and template graphics so they fit (and look great) on just about any screen size or resolution.

Responsive design also gives comfort for the future—an assessment developed today will still be usable in the future, notwithstanding the rapid pace of device technology change.

2. **Blended delivery with mobile support and printing and scanning**
   Blended delivery means that you can create higher- and lower-stakes assessments once and then deliver them on a variety of different platforms such as mobile devices, workstations, and even with printing and scanning. Delivering assessments on mobile devices helps you assess employees "in the field" when it is needed. Delivering paper assessments and scanning in the results is useful for employees without access to technology.

3. **Accessible to employees with disabilities**
   To be compliant with equal opportunity laws, an assessment system must be able to make accommodations for those with disabilities, ensuring Section 508 and similar accessibility compliance. Using systems that cannot make accommodations will expose your organization to unnecessary risk. The Questionmark assessment delivery system is designed and maintained to provide accommodations. For more information, please refer to Questionmark's best practice guide on accessibility.

4. **Results saved at the server "as you go" so IT failures do not lose data**
   Devices and network connections can fail, and it is important that every answer made by a participant is recorded and saved at the server when it is made, so that if there is a problem, answers made so far are not lost. Questionmark's save-as-you-go (SAYG) capability allows saving of every answer as it is made and then resuming the assessment later with the answers recovered. This means that even if hardware or connections are not robust, you do not lose assessment reliability because you can resume later.



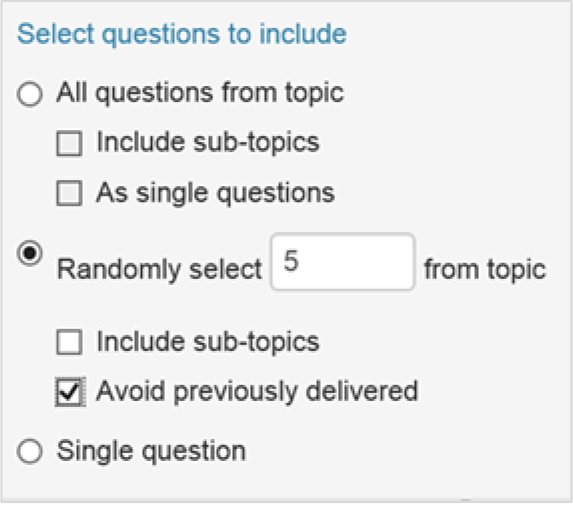
Save as you go

Automatic every 5 minutes

☑ Record results in database

5. **Option to present different questions on a retake**

When someone retakes an assessment after a failed result, it is most valid if they see different questions than the ones they saw in the first attempt. Otherwise, the accuracy of the retake could be impacted by them having seen the questions before.



Select questions to include

○ All questions from topic

□ Include sub-topics

□ As single questions

● Randomly select 5 from topic

□ Include sub-topics

☑ Avoid previously delivered

○ Single question

An effective way to do this is to use the Questionmark capability to "Avoid previously delivered," as seen to the right, when selecting questions at random from a topic. Where there are sufficient items in the bank, this will present different randomly selected ones on subsequent attempts by the same participant.

6. **Forced time period before a retake**

In a similar vein, it is good practice to require a gap between retakes to encourage a participant to restudy before a second attempt. This removes the risk that someone will retake the assessment immediately and hope to get lucky on the next attempt. Questionmark allows you to set a period of one or more days before a retake is allowed.

7. **Option to require monitor to confirm participant identity**

Considering security within assessment delivery, there are three main ways in which people can cheat in taking assessments:

- Identity fraud—an employee might ask a colleague to take the test instead of him/her.
- Content theft—questions are circulated from one participant to another; e.g., someone copies an exam and shares it with a colleague to help them
- Cheating—an employee has a friend sitting with them or uses Internet searches to help answer questions

All of these seriously compromise test validity. The most common method to defeat identity fraud is to have a monitor (also called a proctor or invigilator) identify the person and sign them in for a test. This is easy to set up in Questionmark, and such a monitor—very often for workplace assessments—would be the employee's manager or someone from the HR or compliance department.

8. **Secure browser to provide exam integrity**
   There are many ways to reduce content theft and cheating, including reducing the motivation to cheat by allowing retakes on failure and by making it less likely that people will rationalize to themselves that it is okay to cheat by ensuring that the test is seen to be fair.

   One technical measure to reduce content theft is to use Questionmark Secure, which is a lock-down browser that makes it harder to copy or print screens while taking a test or exam. Using a lock-down browser will also make cheating harder, as it prevents a participant from browsing the internet and using chat sessions to seek advice from others. Questionmark Secure is available for the PC, Mac, and iPad.

9. **Launch and track assessments from other systems of record**
   If your participants are already signed into your organization's portal or learning management system or other system, you typically want to launch the assessment from that system for ease of access and to aid identity verification. You may also want to return summary results returned to that system for tracking. Easy integrations are possible using industry standards (including AICC, SCORM and IMS LTI), single sign-on (SSO) and Questionmark's web services API.

Analyzing results

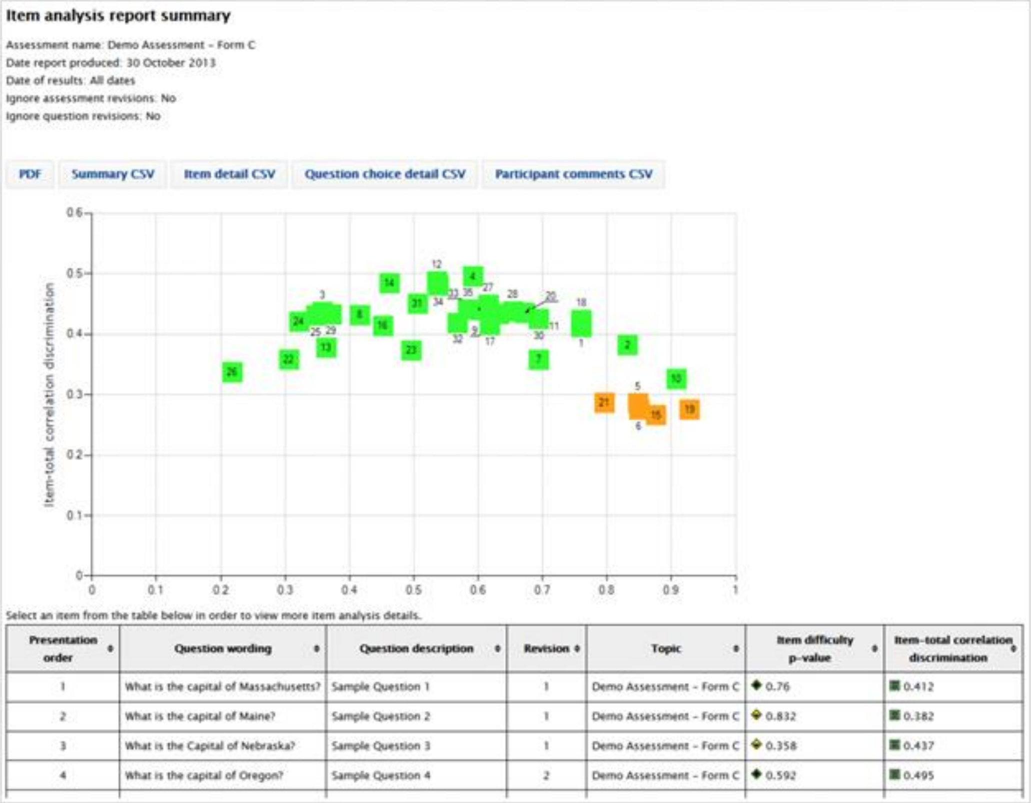
Good reporting is only useful if authoring and delivery have been valid and reliable, but reports must be accurate and clear to be trusted. Some key capabilities required to analyze assessment results are:

1. **Accurate, robust and tamper-proof results storage**
   For trustable assessments, you need participant responses stored accurately even in the event of system downtime or technical issues, and you need to ensure that results cannot be tampered with or lost. Questionmark has over 25 years of experience in developing assessment software and has implemented a highly robust architecture backed up by solid quality assurance practices to ensure that results are stored safely. This is our software's key mission.

   One main way in which Questionmark ensures reliable results storage is that there are two independent databases for results—a transactional database and a results data warehouse. When someone takes an assessment, results are stored immediately in the transactional database; they are then copied into the results warehouse for analysis.

2. **Differential security on results so only the right people see them**
   It is important that assessment results are only available to those who have a need to know. If assessment results are shared inappropriately, this will reduce face validity. In addition, if the results include questions or answers, this could expose questions to future test-takers. However, Questionmark has the capability to restrict assessment results via differential security.

3. **Item analysis report to weed out poor items**
   You need to run item analysis both at the trial stage and when questions are in production. Item analysis looks at the difficulty of items and how they correlate to test results, and it is easy to use Questionmark item analysis reports to identify poor items to improve or remove them. There are a lot of powerful capabilities in the item analysis report, but one of the most useful is its color coding of items.
   For example, in the screenshot below, four items are flagged in amber as needing further review.
   (see next page)



Item analysis report summary

Assessment name: Demo Assessment - Form C
Date report produced: 30 October 2013
Date of results: All dates
Ignore assessment revisions: No
Ignore question revisions: No

Select an item from the table below in order to view more item analysis details.

Presentation order ⬍	Question wording ⬍	Question description ⬍	Revision ⬍	Topic ⬍	Item difficulty p-value ⬍	Item-total correlation discrimination ⬍
1	What is the capital of Massachusetts?	Sample Question 1	1	Demo Assessment - Form C	◆ 0.76	■ 0.412
2	What is the capital of Maine?	Sample Question 2	1	Demo Assessment - Form C	◆ 0.832	■ 0.382
3	What is the Capital of Nebraska?	Sample Question 3	1	Demo Assessment - Form C	◆ 0.358	■ 0.437
4	What is the capital of Oregon?	Sample Question 4	2	Demo Assessment - Form C	◆ 0.592	■ 0.495

4. **Test analysis report to calculate reliability**
   Questionmark's test analysis report includes a calculation of reliability using a statistical measure called Cronbach's Alpha. This is a direct measurement of assessment reliability, and given that reliability is one of the key requirements for trust, this report will measure and tell you how reliable your assessments are.

5. **Wide range of accurate reports to help professionals make good decisions**
   As well as the item analysis and test analysis reports, Questionmark provides a wide range of accurate reports at the right time to help professionals make good decisions. There are around 30 standard reports and also an OData results feed to allow results to feed into other systems and dashboards.

6. **Results can be anonymous to help make employee surveys valid**
   Some surveys will be more valid if the employee is able to answer anonymously—e.g., employee attitude surveys or surveys on cultural issues relating to regulatory compliance. Questionmark lets you make a survey anonymous so that access to it is controlled by scheduling but results do not show the participant's name or details.

7. **"Big data" potential to make better business decisions**
   Last, but not least, there is huge potential in correlating assessment results with other business data to help make decisions. For instance, you could compare and correlate assessment results with individual business performance or other business metrics. The Questionmark results warehouse and OData feed, which can pass data to SAP and other reporting systems, make this a realistic option. If you want to use assessments as part of your "big data" analysis, they must be of good quality; otherwise your analysis will not be meaningful.

Table of key measures effective for reliability and validity

Measure to aid reliability	Questionmark	Your system?
Planning the assessment		
Use job task analysis surveys to help blueprint assessments	✓	
Organize items in an item bank with topic structure	✓	
Metatag items to associate with task knowledge, skill or ability	✓	
Authoring processes secure to protect against content theft	✓	
Manage translations to aid multilingual assessments	✓	
Authoring items		
Authoring tool subject matter experts can use directly	✓	
20+ question types	✓	
Ability to use relevant stimulus including video, audio and equations	✓	
Use observational items/assessments to help measure performance	✓	
Item version history for legal defensibility	✓	
Easy collaboration for item reviewers to help make items more valid	✓	
Search questions to identify those that need updates	✓	
Retire questions that are no longer valid	✓	
Assembling the assessment		
Rules-based or random selection of questions from item bank	✓	
Random ordering of questions and choices	✓	
Set a pass score including topic pre-requisites	✓	
Topic scoring and feedback	✓	
Pilot and Review		
Easily deliver assessments for trial purposes	✓	
Participants can comment on questions and provide feedback to improve quality	✓	
Use experimental questions in live assessments without impacting scores	✓	
Run useful reports on the results of trial assessments	✓	

Delivery		
Consumer-quality participant user interface	✓	
Blended delivery with mobile support and printing and scanning	✓	
Accessible to employees with disabilities	✓	
Results saved on the server “as you go” so IT failures do not lose data	✓	
Option to present different questions on a retake	✓	
Forced time period before a retake	✓	
Option to require monitor to confirm participant identity	✓	
Secure browser to provide exam integrity	✓	
Launch and track assessments from other systems of record	✓	
Analyze results		
Accurate, robust and tamper-proof results storage	✓	
Differential security for results so only the right people see them	✓	
Item analysis report to weed out poor items	✓	
Test analysis report to calculate reliability	✓	
Wide range of accurate reports to help professionals make good decisions	✓	
Results can be anonymous to help make employee surveys valid	✓	
“Big data” potential to make better business decisions	✓	

5. Conclusion

In this white paper, we have seen the value of trustable assessment results. If you are using assessments to make decisions about people, you need results you can trust. Trustable results require assessments that are both valid (measure what you are looking for them to measure) and reliable (consistent in that measurement).

If a doctor orders a blood test, he or she expects that the test will measure what it is designed to measure and will be consistent in what it measures. In a similar way, if you assess someone's knowledge or skill or check the competence of someone for health and safety or regulatory compliance purposes, you want to be sure the assessment is designed correctly and runs consistently.

For assessments to be valid and reliable, it is necessary that you follow structured processes at each step from planning through authoring to delivery and reporting. We have shown a wide range of capabilities provided by Questionmark to help make your assessments valid and reliable. It is important that you follow appropriate processes when planning the assessment, authoring items, assembling the assessment, trialing it, and then when delivering and reporting on it. We have shared over 30 key capabilities of Questionmark that help you do this.

We hope this white paper will be useful to you if you are using Questionmark or if you are considering purchasing Questionmark but need to understand its value compared to other software you have. The authors welcome feedback, questions or suggestions to improve this white paper—please direct them to john@questionmark.com.

Using Questionmark software to organize your assessments makes it easier to create valid and reliable, and therefore trustable, assessments using simpler tools. If you use assessment results to make business decisions, trustable results bring significant benefits."

Appendix – further reading

Here are some useful resources to understand and help make trustable assessments.

Best practice guidance from Questionmark

Questionmark publishes several other white papers giving best practice guidance, available from http://www.questionmark.com/resources/whitepapers. The following white papers are useful for follow-up reading:

- Assessments Through the Learning Process
- Five Steps to Better Tests
- Defensible and Legal Certainty for Tests and Exams

The Questionmark blog at http://questionmark.com/resources/blog includes many useful articles on assessment and provides good practice advice.

Books on trustable assessments

Two useful books on assessment are:

- *Criterion-Referenced Test Development: Technical and Legal Guidelines for Corporate Training* by Sharon Shrock and William Coscarelli
- *Tests that Work: Designing and Delivering Fair and Practical Measurement Tools in the Workplace* by Odin Westgaard

Relevant assessment standards

The following organizational or industry standards are useful to understand:

- ISO 10667: Assessment service delivery—Procedures and methods to assess people in work and organizational settings
- ISO 23988: A code of practice for the use of information technology (IT) in the delivery of assessments
- ISO 17024: Conformity assessment—General requirements for bodies operating certification of persons
- Standards for Educational and Psychological Assessments (from the American Psychological Association)
- International Test Commission guidelines available at www.intestcom.org/guidelines/

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

Questionmark
35 Nutmeg Drive, Suite 330
Trumbull, CT 06611
United of States of America
Tel: (800) 863-3950
Fax: (800) 339-3944
info@questionmark.com

Questionmark
Moor Place, 1 Fore Street
London EC2Y 9DT
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.

questionmark.com