WHITE PAPER

Defensibility and Legal Certainty for Tests and Exams

A Best Practice Guide

This best practice guide is for organisations that use tests and exams (summative assessments) and want to protect themselves from possible legal challenges. This guide helps explain how to create tests and exams with the objective of achieving legal defensibility and legal certainty.

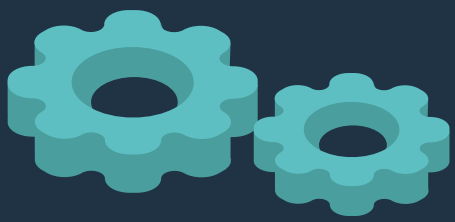
This guide primarily focuses on legal defensibility and legal certainty in a European context, but the best practice suggestions are applicable worldwide.

This guide is NOT legal advice; it should be regarded as general information only and you should obtain your own legal advice for your own circumstances.

Authors: John Kleeman

with

Eric Shepherd
Jamie Armstrong
Sonata Ožemblauskaite



www.questionmark.com
sales@questionmark.com

USA
35 Nutmeg Drive, Suite 330
Trumbull CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
Moor Place, 1 Fore Street
London EC2Y 9DT
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+ 49 (0) 89 220 61272

Contents

www.questionmark.com
sales@questionmark.com

USA
35 Nutmeg Drive, Suite 330
Trumbull CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
Moor Place, 1 Fore Street
London EC2Y 9DT
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+ 49 (0) 89 220 61272

Contents

1. Introduction

1.1 Welcome to this best practice guide

We are all familiar with the concept of a chain of custody for evidence in a criminal case. If the prosecution seeks to provide evidence to a court of an object found at a crime scene, they will carefully document its provenance and what has happened to it over time, to show that the object offered as evidence at court is the object recovered from the crime scene.

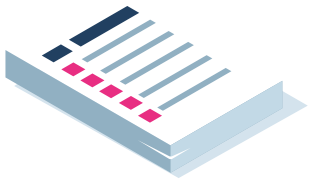
There is a useful analogy between this concept and defensibility and legal certainty in tests and exams. Assessments have a "purpose" or a "goal", for example, the need to check a person's competence before allowing them to perform a job task. It is important that an assessment program defines its purpose clearly, ensures that this purpose is then enshrined in the design of the test or exam, and checks that the assessment and delivery is consistent with the defined purpose. Essentially, there should be a chain from the purpose to design to delivery to decision, which makes the end decision defensible. If you follow that chain, your assessments may be defensible and legally certain; if that chain has breaks or gaps, then your assessments are likely to become less certain and more legally vulnerable.

This best practice guide is for organizations creating and administering tests and exam programs, that are concerned about ensuring defensibility and certainty. The guide describes the principles and key steps to make assessments that are defensible and that provide legal certainty, and which are less likely to be successfully challenged in courts.

The guide focuses primarily on assessments used in the workplace and in certification, however, it will also have some value to those working in other sectors. Given the more litigious culture of the United States (US), there is a well-established body of literature on defensibility in that jurisdiction, but less in the way of reported cases elsewhere. Legal certainty ("Rechtssicherheit" in German) is an important issue in many European countries, but less widely discussed in the US.

The guide explores defensibility and legal certainty in the area of assessments from a broadly European perspective, nevertheless, it will also be relevant in other regions.

Building an assessment program that is defensible, and which encourages legal certainty is not just helpful in protecting from legal challenges. Such assessment programs are likely to be of better quality – giving a better indication of test-taker competence and being more trustworthy for stakeholders to rely on for decision-making. This guide is intended to be helpful to all those working with assessments who seek to make good quality tests and exams.



1.2 The value of assessments

Assessments are used for many reasons, including formative, diagnostic, needs, reaction and summative. This guide focuses on assessments which are used to make decisions about people and that could have a significant or legal impact on them. Such assessments are usually **summative assessments**, i.e. a test or exam, and usually quantitative, with the primary purpose of giving a definitive grade and/or making a judgement about the test-taker's achievement or competence.

There are hundreds of different ways in which summative assessments are used. Here are some common examples:

- Health and safety tests – to check if personnel can work safely before allowing them into facilities or to participate in activities with safety risks;
- Competence tests for regulatory compliance purposes, where companies in finance, pharmaceuticals and other regulated industries test their workforce's understanding of rules and procedures to mitigate regulatory risk;
- Sales or technical channel verification tests, where an organization checks the competence of its sales or technical teams at partners and resellers;
- Promotion tests – to determine who gets promoted within an organization;
- Post-course tests – to determine passing or failing a course;
- Certification exams which give certification and accreditation of skills;
- Pre-employment screening tests – to determine who is advanced to the next stage of a recruiting process;
- Recruitment tests which measure competence or skills to help choose the right applicant for a job role.

These and other summative assessments provide great value to participants, organizations and society, but because assessments of this nature impact people, it is important that they are prepared and delivered professionally and in a way that reduces the chance of dispute.

1.3 Defensibility

Defensibility, in the context of assessments, concerns the ability of a testing organisation to withstand legal challenges. These legal challenges may come from individuals or groups who claim that the organisation itself, the processes followed (e.g., administration, scoring, setting pass scores, etc.), or the outcomes of the testing (e.g., a person is certified or not) are not legally valid. Essentially, defensibility has to do with the question: ***"Are the assessment results, and more generally the testing program, defensible in a court of law?"***.

Short summaries of some relevant law cases are included in this best practice guide to illustrate real-life instances where defensibility and legal certainty were tested by courts. Here is one such case, which arose in the context of alleged discrimination against a group of test-takers.

Case 1. Judicial review of exam for bias

The Royal College of General Practitioners (RCGP) in the United Kingdom (UK) conducts a Clinical Skills Assessment, an exam where a doctor is presented with a series of cases, in which an actor role-plays a patient with symptoms; the doctor must examine the patient and then assess and communicate a diagnosis. It is necessary to pass the exam to practice as a general practitioner doctor in the UK.

In 2014, a legal action was initiated by a group that argued the pass rates for this exam for candidates identifying as "white" were higher than those of other races. The group alleged that the exam was therefore biased against non-white people (including those coming from other countries) and should be considered unlawful – this went to a judicial review in the English courts.

The judge examined the matter in depth and determined that although the statistics for pass rates were not equal between different racial groups, the skills being tested were genuinely needed by a general practitioner doctor and the testing methodology was appropriate. The judge commented that "No better means of testing those skills has yet been devised than the Clinical Skills Assessment." Accordingly, the judge ruled that the RCGP was using the exam in a way that was proportionate to its legitimate aim of ensuring patient safety and so was not discriminatory.

However, although the claim of the group was rejected, the judge suggested that the RCGP had duties to encourage equality of opportunity, and that if it did not take measures to do so in future, subsequent claims of bias might succeed. After the case, the RCGP and the group that initiated the case met and agreed various training measures to help people (including candidates originating from outside the UK) learn the skills needed to pass the exam.

The key to successfully defending this exam was showing that the skills tested were relevant to and aligned with the job task for which the exam was being used.

Ensuring that assessments are defensible involves two key steps:

1 You need to ensure that the assessments are:	◆ **Valid, in that the assessment matches its purpose;** ◆ Reliable, in that the results are within an acceptable range of measurement error; ◆ Fair, in that the results are objective and do not advantage or disadvantage individuals based on irrelevant characteristics
2 You need to ensure you have evidence and documentation available to demonstrate the above, in case of a formal or even legal challenge	

1.4 Legal certainty

Certainty is a legal principle of broad, general application, including within the European Union (EU). It means that the law (or other rules) must be certain, in that the law is clear and precise, and its legal implications foreseeable. If there is legal certainty, people should understand how to conduct themselves in accordance with the law. This contrasts with legal indeterminacy, where the law is unclear and may require a court's ruling to determine what it means.

Lack of legal certainty can provide grounds to challenge assessment results. Here are some circumstances where challenges based upon legal certainty may arise:

1. Many organisations have rules for how they administer assessments or make decisions based on the results of assessments. A test-taker might claim that the organisation has not followed its own rules. Alternatively, there may be an ambiguity or gap in the rules or guidelines and/or the testing organisation seeks to interpret them in some way that is different to test-taker's initial understanding based on information provided or reasonable inference.

Case 2. Claiming lack of legal certainty due to failure to follow rules

A candidate for the qualifying examination for the European Patent Office scored very close to the pass mark. On one of the papers, he was given a score of 44 by one examiner and 45 by another examiner. Under the normal rules, the candidate's score would have been averaged to 44.5 and rounded up to 45, and with a score of 45, he would have passed the set of exams. However, the Examination Board decided to appoint a third examiner, who gave a score of 43. Including a third examiner was not part of the formal rules. However, the third examiner's score was considered as part of the decision-making process and the candidate was failed.

The judgement of the European Patent Office Board of Appeal was that the Examination Board had exceeded their powers. It advised that one of the fundamental principles of a fair procedure is legal certainty, i.e. the right of parties to know the basic rules of the procedure in advance. The appointment of a third examiner without any basis in the regulations was deemed a substantial procedural violation.

The exam result of the candidate in this case was remitted to the Examination Board for re-review.

The lesson learned here is that if an organization has rules, it must follow them. This applies particularly to public bodies, but also likely to any organization that publishes its formal rules.

2. In some cases, an organization's actions are constrained by law. For example, a government agency may only be able to do things which a specific law permits, and it is possible in some contexts that the law may prohibit the use of assessments. As an example of this, there has been some debate within certain German universities regarding whether laws and regulations that permit written exams also permit computerized ones[1]

1 See Rechtliche Aspekte von E-Assessments an Hochschulen (Legal aspects of E-Assessments in Universities) available at **http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=42871** (in German)

3. Legal certainty issues can also arise if the exam process goes awry. For example, someone might claim that their answers have been swapped with those of another test-taker, that there has been a mistake in adding up the score or that the exam was unfair because the user interface was confusing, e.g. they unintentionally pressed to submit their answers and finish the exam before actually intending to do so

Case 3. Claiming lack of legal certainty relating to assessments

In 2015, the European Union Civil Service tribunal gave a judgment in response to a claim from a Hungarian national who was denied a job at the European Union Agency for Fundamental Rights. One of her arguments was that she was not informed in advance of how the tests were marked and of the scores required to pass. The candidate claimed that this breached the principles of legal certainty and transparency. She argued that EU rules required candidates to be advised of how tests were marked and that there were other procedural irregularities in the process of rejecting her for the job role.

The tribunal ruled that the agency had followed its regulations appropriately. For some types of recruitment, it was necessary to inform candidates in advance how tests were marked, but this was not necessary in the present case. In summary, the procedural complaints were not valid and so the claim was rejected.

The EU has official documents which define recruitment procedures and the ruling suggests that if the agency had genuinely violated these rules, the applicant would have had a stronger case. This underscores the importance for a public body of having appropriate rules, following these correctly and documenting that all procedures have been followed.

As well as issues relating to defensibility, this best practice guide also includes various good practice suggestions for encouraging legal certainty, such as ensuring that there is robust evidence in place of what happened while the test or exam was taken.

1.5 ISO 10667

In any field of practice, it is helpful to refer to accepted best practice standards for several reasons. Firstly, these often reflect the considered and consensus experience of experts in that field. Secondly, when defending practices, it is useful to say that you are following best practices. Perhaps most importantly, a failure to follow a standard or good practice exposes an organisation to a heightened risk of criticism, since it is difficult to argue against following a good practice standard.

One very respected set of standards in the assessment field is the Standards for Educational and Psychological Testing[2], developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA) and the National Council on Measurement in Education (NCME). This is widely used by high stakes testing programmes.

In this best practice guide, we refer to an International Standards Organisation standard, ISO 10667. ISO 10667 is titled "Assessment service delivery - Procedures and methods to assess people in work and organizational settings" and is a high-level guide setting out good practice standards in workplace assessments. ISO 10667 has two parts (ISO 10667-1:2011 and ISO 10667-2:2011); part one is aimed at the client ordering an assessment service and part two is aimed at the service provider who is supplying assessment services to the client[3].

2. See **www.apa.org/science/programs/testing/standards.aspx**

3. See **www.iso.org/standard/56441.html** and **www.iso.org/standard/56436.html** for more information on ISO 10667.

2. Ten Key Considerations

This section introduces ten key concepts and issues that you need to be aware of in creating and delivering tests and exams that are defensible and encourage legal certainty.

2.1 Documentation

The first place to start in any defensible assessment project is with documentation. You need to document every step of the process, starting with a test specification, continuing with documenting how the test is authored, how the cut or pass score is set and ending with documentation of delivery, reporting, dispute resolution procedures and governance. As a legal handbook for trainers puts very eloquently:

> ***"In any legal dispute … the agency representative or civil jurors must ultimately determine whose recollections are most credible. The existence of effective, accurate and consistent documentation helps immeasurably. Conversely the absence of documentation can destroy even the best technical defense."*** [4]

It is important to keep records of the development of your tests and ensure that these records are updated so that they accurately reflect what you are doing within your programme. These records will be powerful evidence in the event of any dispute. Further details on what ought to be documented are provided later in this best practice guide. It would also be worth referring to the documents mentioned in the appendix of this best practice guide for more detailed guidance.

2.2 Consistent procedures

Testing is more a process than a project. Tests are typically created and then updated over time. It's important that procedures are consistent over time. For example, a question added into the test after its initial development should go through similar procedures as those for a question when the test was first developed. And the administration process must be consistent for all test-takers.

4 Patricia Eyres, "Legal handbook for trainers, speakers and consultants" 1997, McGraw Hill

If you adopt an ad hoc approach to test design, fail to consistently deliver the test with the same administration procedures, or otherwise are inconsistent in your procedures, you are exposing yourself to an increased risk of successful legal challenge.

2.3 Validity

Let's move on to the three generally accepted principles of good test design – the first of which is validity. The other principles of reliability and fairness are addressed in the immediately following sections.

> ***One of the most important characteristics of any test is its validity. ISO 10667-1:2011 defines validity as the "degree to which the interpretation and use of assessment scores are consistent with the proposed purposes of the assessment and are supported by accumulated evidence and theory".***

Essentially, validity is how well the assessment matches its purpose. For example, if you have a test that seeks to measure a skill, you need to check how well the test scores actually measure that skill, as well as whether there is associated evidence to indicate the validity of the assessment. One of the most important aspects of validity is content validity. This means whether assessment content and composition are appropriate, given what is being measured, e.g. does the test cover knowledge/skills required to perform a job.

You need validity evidence in order to rely on an assessment to make decisions, and if you make decisions using an assessment without having any evidence that the assessment can validly be used for this purpose, you will find it hard

to defend these decisions. You should always consider validity before using the assessment to make decisions, not just if you are challenged. As the US Equal Opportunity Employment Commission wisely says:

"Validation studies begun on the eve of litigation have seldom been found to be adequate." [5]

We refer to validity frequently in this guide, including with respect to job analysis in 2.6 below.

5 See **http://uniformguidelines.com/qandaprint.html**

2.4 Reliability (also called precision)

As well as being valid, tests also need to be reliable (or precise). The reliability of an assessment is defined by ISO 10667-1:2011 as the "degree to which scores are free from measurement error variance, i.e. a range of expected measurement error".

One way of encouraging reliability is to have longer tests – all other things being equal, more questions mean that you can be surer of the result, but this obviously needs to be balanced with practicality. The best practice recommendations in this guide seek to help you to produce valid and reliable tests.

2.5 Fairness (also called equity)

Probably the biggest cause of legal disputes over assessments is whether they are fair or not. ISO 10667-1:2011 defines equity as the "principle that every assessment participant should be assessed using procedures that are fair and, as far as possible, free from subjectivity that would make assessment results less accurate". A significant part of fairness/equity is that a test should not advantage or disadvantage individuals because of characteristics irrelevant to the competence or skill being measured.

One example of where fairness is particularly important is when someone has a disability. Many countries have legal requirements to make reasonable accommodations for people with disabilities, and if it is possible for someone to do a job with a disability provided such adjustments are made, it is important that the test is fair in respect of giving any similar reasonable accommodations for people with disabilities. Another example is language; you should consider if it is fair to ask questions in a language that the test-taker is not fluent in. Later sections of this best practice guide will cover aspects of fairness in more detail.

2.6 Job and task analysis

The type of skills and competence needed for a job change over time. Job and task analysis are techniques used to analyse a job and identify the key tasks performed and the skills and competences needed. A key component of validity for tests in the workplace or that are used to recruit is how well the test aligns with the job that the test-taker is doing or seeking to do. It is important that you can demonstrate with evidence that a test used for someone for a job role is aligned with that job role.

For example, if you are using a test to help choose which police officers get promoted into senior roles, it is important that the test is aligned with and representative of the job tasks that are important in those senior roles. Often a job task analysis is performed to provide this job analysis, as explained in section 3.2.

For assessments used in the workplace, a job analysis is a bit like the foundations of a house – if you do not do it, the whole house will be unstable. And if you use a test for a job without having some kind of analysis of job skills, it is likely to be very hard to prove and defend that the test is actually appropriate to measure someone's competence and skills for that job.

Case 4. Example of a legal case illustrating job task alignment

In the United Kingdom, barristers (advocates) are required to pass an exam before they are able to practice. The exam has several modules and one module is to research a legal matter and write an "opinion" (a legal explanation) on it under time constraints. The regulations for the exam state that a participant only has two attempts at each module, bar certain extenuating circumstances. And if there is a failure on both attempts at any module, then the candidate is considered "Not Competent" and must retake the entire course and all exams.

A prospective barrister passed all elements of the exam except this module, which he failed twice, and so was told that he could not practice and would need to retake the course and exams. He felt that this was unfair, partly as the cost was prohibitive and partly because he felt that he had demonstrated sufficient skills for the job. A legal action on the matter was raised in the English courts in 2015.

In its judgement, the court rejected the claim due largely to considering that the assessment had job task alignment. The judge ruled sufficient evidence existed to justify that being able to write

an opinion under time constraints was part of the required skills for a barrister. Therefore, it was reasonable for the examining authority to require that candidates pass this module, as a matter of protection of the public from someone performing the job without the necessary skills.

Key to the judge's decision was that the requirement for this skill was well documented and needed to do the job, so that failing a candidate for not having the skill was fair. A second key reason why the case did not succeed was that the exam regulations clearly identified that repeated failure of any module would result in failure of the whole exam. .

2.7 Setting the cut score (or pass score)

A cut (or pass) score is the score that a test-taker needs to achieve to pass a test. Setting this defensibly is critical because if you do not, then there is increased possibility that a test-taker may be able to successfully claim that he or she was failed unfairly. So how should you set a cut score? Should everyone over 70% pass, and everyone under fail? Or is it better to set a higher cut score of 75%, or 80%?

There are two types of test commonly used: norm referenced and criterion referenced:

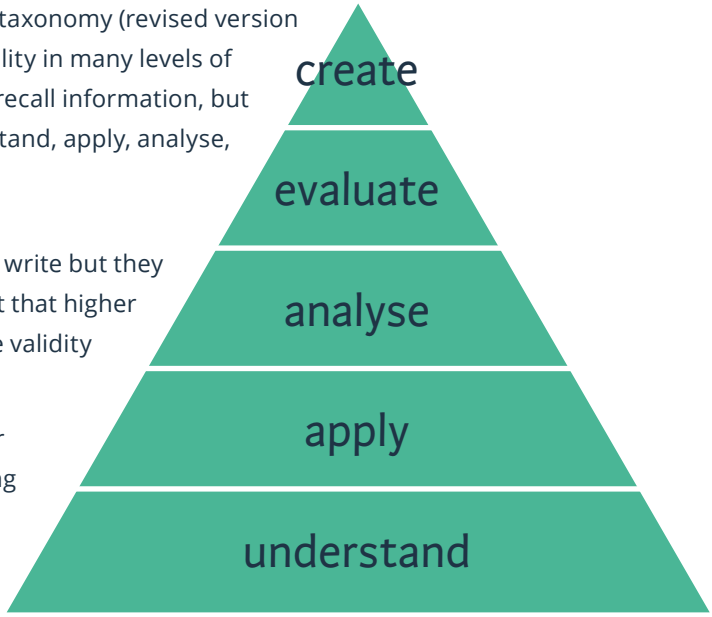
- In a norm referenced test, the test-taker is compared to others in the same group. This is typically used for educational tests or tests where only a certain number of test-takers can be selected. The pass/fail decision can only be determined after the assessment and depends on how others in the test score
- In a criterion referenced test, the test-taker is compared to a benchmark. This is typically used for certification, pre-employment tests, job readiness assessments and compliance or safety tests. Typically pass/fail can be determined before the assessment

With criterion referenced tests you need to set a cut score that distinguishes the minimally competent from those who are not competent. If a test-taker passes the test, he or she is judged as competent. If a test-taker fails the test, he or she is judged as not competent and may have to retake the test or suffer other consequences

As explained later in section 4.2, it is important that you have evidence to reasonably justify that the cut score does genuinely distinguish the minimally competent from those who are not competent. You should not just choose a score of 60%, 70% or 80% arbitrarily, but instead you should work out the cut score based on the difficulty of questions and what you are measuring.

2.8 Questions focused on more than just knowledge recall

It is useful to think of questions in terms of the Blooms taxonomy (revised version shown right). Most real-world jobs and skills require ability in many levels of the taxonomy. People need to be able to remember or recall information, but that is only a part of the skill – they also need to understand, apply, analyse, evaluate and create.

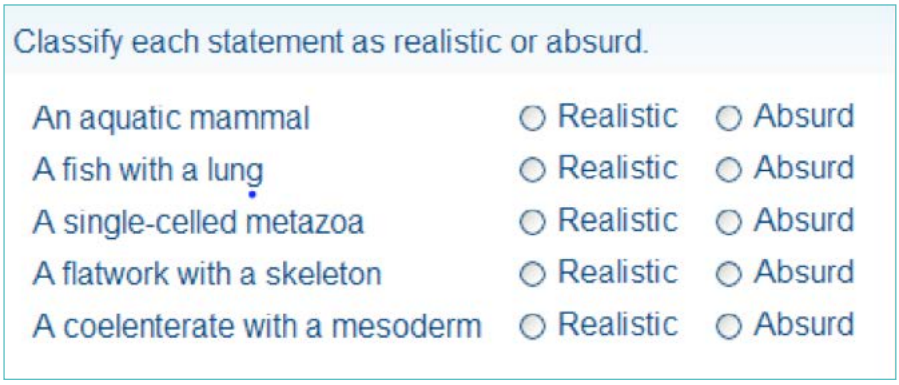


Questions which test remember/recall skills are easy to write but they only measure knowledge. For most tests, it is important that higher level skills are measured as well. And if they are not, the validity of the test is threatened.

It is important to consider approaches on testing higher than knowledge at the test specification and item writing stage. One approach to measuring higher level thinking skills is to make the question stimulus emulate the performance environment that a person might encounter on the job, by presenting scenarios, video, sounds, observational assessments and/or simulations. Additionally, higher order skills can be measured by using well-written objective questions which require thought to answer, see for example the question in section 2.9 below. It is also useful to use situational judgement questions (see section 4.1).

2.9 Consider other question types as well as multiple choice

When selecting question types to use you must consider the cognitive and/or motor mechanical skills required to answer the question and whether these skills will be required by someone to successfully perform their tasks on the job. For instance, a drag and drop question type might disadvantage people that would be perfectly able to perform the tasks required for a job. There is good psychometric evidence that multiple choice tests can assess well; however, particularly in Europe, multiple choice questions sometimes get a "bad press". There is concern that the answers to multiple choice questions can be guessed, and also it is harder to write questions that test higher order skills in a multiple choice format.



As you design your test, you may want to consider including enhanced stimulus and a variety of question types (e.g. matching, fill-in-blanks, etc.) to reduce the possibility of error in measurement and enhance stakeholder satisfaction. For example, the question to the right tests higher order biology skills and only has a 3% chance of someone guessing the right answer vs a 25% chance in a 4-choice multiple choice question.

You can deliver valid, reliable and fair tests when only using multiple choice questions, but in some contexts, your stakeholders may be happier to see other question types used.

2.10 Robust and secure test delivery process

A critical part of the chain of evidence is to be able to show that the test delivery process is robust, that the scores are based on answers genuinely given by the test-taker and that there has been no tampering or mistakes. This requires that the software used to deliver the test is reliable and dependably records evidence including the answers entered by the test-taker and how the score is calculated. It also means that there is good security so that you have evidence that the right person took the test and that risks to the integrity of the test have been mitigated.

In this guide, section 5 focuses largely on robust test processes.

3. Best practice: Planning the assessment

3.1 Project organisation and planning

Sections 3, 4 and 5 of this guide cover some recommended practical steps to take when seeking to deliver assessments that provide defensibility and legal certainty. As with many things, preparing well is the key to success. We are now going to look at the key preparation steps that contribute to defensibility and legal certainty.

Engage stakeholders and agree to the scope and governance

Defining the scope of the assessment programme will help engage stakeholders. A scope helps define the purpose of the programme and test(s) and guides the systems and processes used in the development, delivery, quality control, score reporting and appeals of the test which are discussed below.

You also need to document how the assessment programme will be governed and consider how many people should be on the governance board, and how often and what should trigger the governance board to meet. You also need to cover appropriate financial and other resourcing for the test programme and how will fees to take the assessment, if any, be determined.

Establish an appropriate and controlled document repository

As we discussed in section 2.1, a key to defensibility is to maintain documentation about the assessment process. You should ensure that documents relating to the assessment development and usage are stored in an appropriate document repository (often Microsoft SharePoint or another similar system is used). It is useful if this can keep track of date/time and versions to provide credibility if the documentation is needed to defend your assessments against legal challenge.

Document the competence of the people involved in the assessment programme

ISO 10667 requires that organisations delivering assessments ensure that people with a role in the assessment process have the necessary competence or are trained to have it. It is helpful to keep a list of people involved in the assessment creation process and documentation of their competence or training – both in assessment methodology and in the subject areas being assessed. This can be helpful evidence if an assessment is challenged.

Set up appropriate roles and permissions in your assessment management system

To be able to trust the results of online assessments, you need to control who has access to authoring, administering test delivery and results. Failure to do this might lead to information leaks, which consequently would increase the risks of cheating and data breaches. You should set up appropriate roles and permissions, and allocate people to them, as well as ensuring that there is a process to keep this up to date and reviewed as personnel change.

Document repeatable procedures

An assessment is not usually a one-time event. The steps to author, review, deliver, monitor and report on an assessment are repeated over time. The only way to ensure that these steps are completed consistently over time is to ensure that the processes behind the steps are documented. You should define the processes for test specification creation, monitoring and review. Also, define the process for item authoring, assessment assembly and pilot, and for assessment delivery, as well as how you will monitor and review the assessments. Set up an approval procedure for such processes together with a review cycle to ensure that the documentation stays up to date, and that processes are reviewed based on experience.

Ensure that your organisation owns or has rights to use all material used in the assessment

If you are using any external authors, translators or reviewers to help create assessments, ensure that legally reviewed agreements are in place to give you ownership of or the right to use the output of their work under intellectual property laws. Also, if you are using any external material or media within the assessment, you should ensure that you have the legal right to use that. Ensuring that you retain appropriate documentation for this will protect you against legal challenges in this area.

If you plan to translate an assessment, do it professionally

Translating questions and assessments accurately so that they measure the same construct fairly in each language requires a professional approach. There have been lawsuits attacking exams because a translation was of poor quality[6]. If you need to translate an assessment, please follow good practices. It is usually sensible to commission a professional expert translation company that specialises in translating assessments and to export text to their systems and re-import it. A useful resource is a recorded webinar "Translating test items: More to it than meets the eye![7]".

6 See **https://timesofindia.indiatimes.com/india/sc-stays-madras-hc-ruling-on-grace-marks-for-neet-in-tamil/articleshow/65076449.cms** for an example in India where a poor translation into Tamil caused a legal challenge.

7 Available at **http://pages.questionmark.com/WC-ENUS-WebinarRecording-Translating-201810_LP-Registration.html**

3.2 Specifying the assessment

A test specification (sometimes called a test content outline or test blueprint) is a master document that defines the purpose of the test and what will be assessed. You should ensure that you have a specification in place before you author questions or progress other steps. Here is some guidance on creating a specification.

Start with the assessment's purpose or rationale

Part of defining the validity of the assessment is document how the use of assessment scores is consistent with the proposed purpose (or rationale / objective) of the assessment, for example to be able to demonstrate that if someone passes a test that this demonstrates a particular competence.

It is important to document the purpose of the assessment and ensure that everything else related to the assessment follows this. For example, the purpose of an assessment might be to check that someone knows and understands the safety rules of their work environment before being considered qualified to work in it; or the purpose could be to ensure that someone has the necessary competence to be allowed to sell a regulated product or service to the public.

Some follow-up questions should be asked once the purpose has been defined. These will help clarify the definition of purpose and allow you to be sure that you have clarified ambiguities before you start work:

- Who is allowed to take the assessment? Are there any pre-requisites?
- What construct or domain is being measured?
- What are the consequences of passing the assessment?
- What are the consequences of failing the assessment?
- Are the scores permitted to be used for any other purposes?
- Who has access to the scores and results and why?
- Are multiple attempts allowed, and is there a minimum time period needed between retakes?
- Is this a norm referenced or criterion referenced test?

- Should the test be open book (test takers are allowed to consult materials during the test) or closed book (they are not)?
- When will the results of the test be disclosed to the test-taker?
- If a test-taker disputes their results, how does the dispute resolution process work?
- How will passing the test be recognised (certificate, badge, etc)?

Identify the audience for the test and any diversity or fairness issues

You need to document the expected participants for the test and consider:

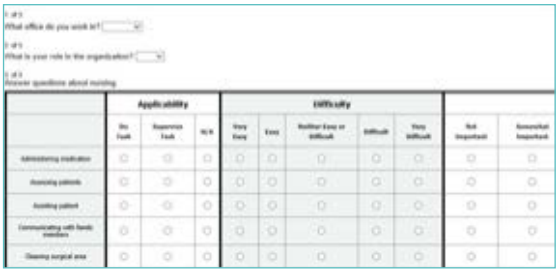
- Is it fair to deliver the test in just one language, or do you need to translate the test?
- What cultural issues are there in within likely test-takers, to be sure not to write questions with cultural references that some will not understand?
- Can you assume a high degree of computer literacy in test-takers or will they need to be guided through any computer use as part of the test? (For example, a test that involves a lot of typing might be unfair to test-takers who are not used to typing)
- What accessibility or special needs issues will you need to take into consideration?
- Are there any other adverse impact or bias issues you need to consider (for example, to ensure that the test is fair to all genders and races)?

Define the retake and result invalidation policies

If a test-taker fails the test, can he/she retake it? It is common to establish a minimum period between retakes (to encourage learning and preparation for the retest). Often, there is a limit to the number of retakes allowed and, if someone is caught cheating, that person is usually prohibited from retaking the test. It is vital for you to define a policy, justify it based on the purpose of an exam and get it approved by stakeholders. If someone is not allowed to retake a test, it could well cause a legal challenge, particularly if that individual's livelihood is at stake, so this is something to define clearly in advance and have well-justified. You also need to define the policy and procedures around invalidating a result, for example if someone is suspected of cheating.

Job task analysis

As we saw in section 2.6, it's critical to do some analysis to provide evidence that a test assesses the skills required to perform the tasks of the job. One technique for job analysis is to conduct a job task analysis (JTA) survey. Often a JTA will ask respondents about the applicability, difficulty, importance and/or frequency of each task, as shown in the fragment shown to the right about nursing skills.



JTA surveys are administered to those performing the tasks of the job and the results are then used to determine the tasks that the test will focus on and thereby the knowledge and skills that must be assessed by the test. JTAs provide evidence of the tasks to be performed, cognitive or motor-mechanical skills required, and provides the anchor, justification, depth, breath and method for the knowledge and skills to be assessed.

A JTA can be good evidence that the test is appropriate for the job. For new job roles, you can use a derivative of JTA whereby subject matter experts predict the "must have", "should have", and "nice to have" knowledge and skills required to perform the projected tasks. If you do not use JTAs, or some other kind of job analysis, you risk that the assessment will not cover the right competencies and tasks, and so it will be invalid and likely indefensible if challenged.

Identify the content areas for the test and the number of questions in each content area

Whether via a JTA or otherwise, you need to define which content areas the test is to cover. For a test evaluating job skills, the content areas will arise from the job analysis. For a placement test, the content areas will derive from the skills needed for the course being placed into. And in other cases, the purpose of the test will drive the content selected.

The number of questions for each topic or objective usually depends on:

- How critical the objective is (the more critical, the more questions);
- How large the domain the objective covers (the larger, the more questions);
- Whether the objective is related to other objectives (the less it is related, the more important it is to ask more questions on it, as knowledge/skill may not be correlated with other objectives).

In their excellent book on criterion-referenced assessments[8], Shrock and Coscarelli suggest the following table as a guide for how many questions to include:

Criticality?	Domain size?	Related?	# questions
Critical	From a large domain	Unrelated	10-20
		Related	10
	From a large domain	Unrelated	5-10
		Related	5
Not critical	From a large domain	Unrelated	6
		Related	4
	From a large domain	Unrelated	2
		Related	1

8 Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training 3rd Edition, Shrock and Coscarelli, 2007, Pfeiffer

Consider whether an observational or other practical element is needed

Some skills cannot easily be measured by a conventional test as they involve practical skills. In such situations, it is common to include a practical test alongside a cognitive test. A useful analogy is learner driver tests – in many countries, there is a theoretical test done online or on paper and a practical test performed in a car. To justify and defend your testing programme, you need to consider and document whether this is required.

One approach here is to use an observational assessment, where an observer (supervisor or instructor) logs in and observes and uses a checklist to rate a test-taker on his/her performance. An example of such an observational assessment is the Objective Structured Clinical Examination (OSCE) used to observe and grade healthcare workers such as doctors and nurses. It is common to use tablets or smartphones for such observational assessments, as they can easily be taken into the workplace. 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.

Document the test specification

It's important to document all of this in a test specification (or blueprint). This describes the content areas (objectives or topics) and the weighting or number of questions in each, the kind of skills within each content area, what kind of questions will be used and more. The test specification will often contain answers to the questions under "purpose" above and other information justifying the decisions taken in test design.

The more you can document at the outset, including the reasoning for decisions taken, the more useful the test specification will be in defending the test later. Sometimes a summary of the test specification is shared with test-takers to help them prepare for the exam and with stakeholders so that they understand the value of the exam.

Set up topics in an item bank

An item bank is a database of questions, usually structured in hierarchical topics and often permitting metatags. An item bank gives:

- An easy management view of all questions and assessments under development;
- Mapping of topics to relevant organisational areas of importance;
- Clear references from questions to topics;
- Use of the same questions in multiple assessments;
- Simple addition of new questions within a topic;
- Easy retiring of questions when they are challenged or expire;
- Search capabilities - for example, identifying questions that need updating when laws have changed or a product has been retired;
- Version history of the assessment project which provides defensibility evidence.

This section has covered many useful preparation steps. The next section looks at the actual authoring process for the test.

4. Best practice: Creating the assessment

4.1 Author questions

The creation process is usefully divided into authoring questions, constructing the assessment and running a pilot.

Here is some guidance on authoring questions.

Set up a collaborative team including subject matter experts

Valid, reliable and fair assessments require a team to create and review them. It is critical to involve subject matter experts in the writing and/or review. Failure to do so will usually make it much harder to defend the assessment results. You should ensure that you have legally reviewed confidentiality agreements in place with your team and regularly train your question authors and reviewers on good practices.

A question style guide is helpful

For consistency, it is very helpful to have a project style guide which sets out rules for question creation. Such a guide might cover:

- Question types and how many choices there should be in questions;
- Glossaries for words that should be used for consistency;
- Guidance on use of acronyms and abbreviations;
- Any cultural guidelines (reminding authors of the potentially international nature of the audience);
- Accessibility guidelines.

A project style guide can contribute to building a defensible assessment that contains consistent items and is developed using repeatable, documented procedures.

Follow good item writing techniques

Poorly written questions are less likely to be legally defensible. In your style guide or elsewhere, you should include common good practices on writing questions (e.g. avoid negatives, avoid grammatical clues in the question stem). There are many resources on good question-writing techniques: a possible one to use is Questionmark's webinar on ***"Item Writing: Tips and Techniques for Writing Good Questions"***.

Set up and follow a consistent review procedure for questions

It is common to set up a review process for questions, where they are written and then reviewed prior to being used in practice. Documenting that all questions go through such a process is extremely helpful both for avoiding errors and for later defence of the questions. Such review should include review of content, language and potential bias.

Case 5. Appeal when a question is wrong

The European Patent Office (EPO) runs examinations for patent lawyers which candidates need to pass to be allowed to represent applicants before the EPO. There is a pre-examination, which must be passed first, and then a main exam consisting of four papers. There is only one sitting of the exams only each year and passing them is an important requirement for a patent lawyer in Europe.

The pre-examination consists of a series of 20 questions, each of which consists of a series of sub-questions that the test-taker must answer true or false. In 2014, a candidate narrowly failed the test and raised an appeal that he/she should have passed because one of the questions was scored wrongly – the correct answer should have been "false" but in fact an answer of "true" was marked as correct.

The candidate's appeal to the Examination Board was rejected, and so the issue was passed to the EPO Disciplinary Board of Appeal (Board of Appeal)which heard the matter and issued a judgment in the case.

The decision of the Board of Appeal was straightforward:

- The candidate was correct that the question had been incorrectly scored;
- Although the examiners had thought the answer to be true, in fact due to the wording of the question, the correct answer was false;
- So, the candidate had answered correctly and should have passed;
- The appeal board ruled that the candidate be considered to have passed the exam

This case emphasizes the need to review questions. It also demonstrates the potential risk of using questions that ask for a true/false answer, since there are often nuances in real-life situations that need to be considered. There should also be an easy-to-access appeals procedure to deal with mistakes in questions.

Retain version and review history for questions

An assessment management system should allow you to record review comments made on questions and track different versions as they are updated. Being able to demonstrate with evidence that questions were reviewed and updated in the light of reviewer comments will be helpful for defensibility in the event of legal challenge.

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" the 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.

Consider situational judgement questions

A situational judgement question presents a dilemma to a test-taker and asks them to choose from a number of possible responses, for example asking them to select the best and worst choice from a series of options. It allows the measurement of judgement that can be very relevant for many job skills. See **https://blog.questionmark.com/tag/situational-judgment** for more on situational judgement questions.

4.2 Construct assessment

Once you have created the questions, you need to put them together in an assessment. Here is some guidance on this.

Select questions to match the test specification

In some cases, you will choose to build "forms", meaning fixed instances of the assessment with specific questions. In other situations, you will choose to build an assessment which selects randomly from the item bank to create a different instance for each test-taker. In either case, it is critical that every test instance matches the defined test specification, and that you can demonstrate with evidence, if required in a court of law, that this is the case. Usually, this is simple to do as you will organise your item bank along topics defined in the test specification, and it will be straightforward to show that the assessment selects appropriately by topic.

Avoid negative marks and usually weight questions equally

It is very common to weight all questions in a test the same, for example, there might be 50 questions and each scores 1 point, thus the test is scored out of 50 points. Usually if you want to have more score weighting associated with a topic area, you would simply use more questions in the topic. However, sometimes you might have a valid reason to give some questions higher weight, for example, a more important or difficult question could score 2 points or higher. If you do this, you need to have very clear justification that relates to the test specification and test purpose, and which can follow through in your cut score definition described below.

It is usually inadvisable to use negative scores (subtracting marks in the test for a poor answer). Subtracting scores needs careful validity justification and use of such negative scores can open up accusations of unfairness – for example, there have been some German legal cases involving disputes on negative marking[10].

Consider pros and cons of randomising question selection

Advantages in randomising question selection include:

- Each test taker gets a different test, which makes cheating harder;
- Because each test is different, it is easier to deliver the test on demand rather than only at fixed times;
- It is possible to allow someone to retake a test and they will see different questions from the first attempt. (In some systems like Questionmark, you can set this to happen automatically so that someone retaking a test gets questions randomly selected from those they did not see in a previous attempt);
- If you need to remove a question, you can simply retire or remove it from the item bank, and the test will continue to select other questions from the topic

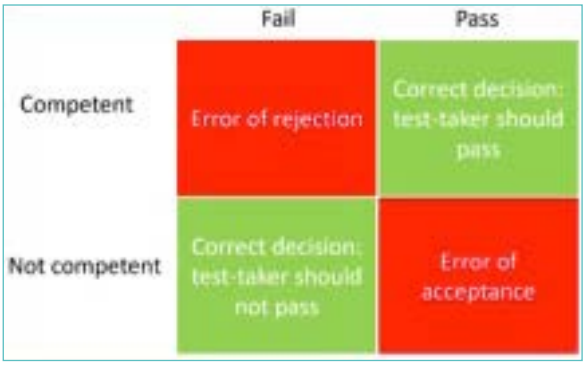
10 See for example **https://openjur.de/u/765071.html** (in German).

However, if you do randomise question selection, you need to consider the risk that someone might get an easier set of questions than another test-taker and so find it easier to pass. A variety of approaches to mitigate this risk are possible. One option that is commonly used is to pilot the questions to determine their difficulty, and so divide them into three categories – easy, moderate and hard; then the test can select an equivalent number of easy, moderate and hard questions in each topic for each test-taker.

Note that although you must justify randomising question selection, it also makes sense to randomise the order of questions and to shuffle the order of choices within a question. These actions make it harder for test-takers to communicate right answers to other test-takers and so reduce cheating.

Set a cut / pass score in a defensible way

Following on from section 2.7, it is vital that you set the cut score defensibly for a criterion referenced test. As shown in the diagram, there are two possible errors when setting the cut score. If you set it too low, then you increase the risk of errors of acceptance, where someone who is not competent passes the test. If you set it too high, then you increase the risk of errors of rejection, where you reject someone who is competent. Errors of rejection are most likely to lead to challenges from test-takers[11].



There are a number of generally accepted procedures for setting the cut score. These do not usually involve statements like "we have always used a cut score of 70%" and so we will keep on doing so. The most accessible method is called the Angoff method (or sometimes the modified Angoff method). This involves getting a group of subject matter experts to assess the probability that a marginally competent test-taker will get each question right. The cut score is then calculated from the sum of these probabilities.[12]

11 See this webinar by the U.S. Coastguard about their approach in this way:
https://www.questionmark.com/content/randomly-designed-tests-how-can-they-be-fair-all

12 For more on the Angoff method, see **https://www.questionmark.com/content/using-angoff-method-set-cut-scores**

4.3 Pilot the assessment

A pilot obtains evidence that helps validate the assessment and allows you to correct issues before you deliver a test in a production environment. Here is some brief guidance on piloting:

Conduct a pilot

You should always run and document a pilot (also called a field test). The nature and size of the pilot will depend on the stakes of the assessment and may be impacted by what is realistic. Ideally the pilot should deliver the assessment to a representative sample of test-takers with enough numbers to permit statistically useful decisions about the draft assessment and its questions.

For an ongoing programme, it is also useful to pilot new questions or new versions of the assessments, either as separate pilot deliveries or by including new questions in production tests as experimental (unscored) questions.

Run item analysis on the pilot results

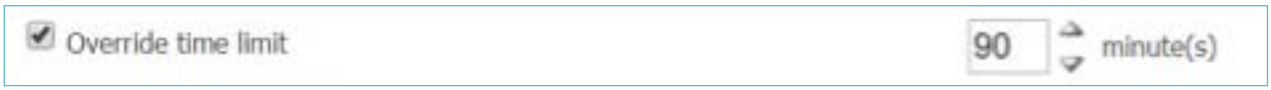
Item analysis is a powerful statistical technique that lets you identify weak questions which you can either remove or improve. See section 6.2 for more on item analysis. If you have enough data, it is also useful to be able to statistically check the reliability of your assessment.

Set a time limit in a defensible way

A pilot provides useful evidence to set the test time limit in a defensible way.

Most tests are “power” tests. They seek to measure knowledge or skills of a test-taker and most people should have enough time to answer most of the questions. For a power test, you can set the time limit based on experience in the pilot. Some tests are “speed” tests, where fast speed is an important part of job requirements. In speed tests, many people will not finish all questions, and you should set the time limit based on the job requirements.

In all tests, you should ensure that the time limit does not start until instructions are given, any demographics are captured and any practice items are completed. It is common to override the time limit and give extra time for people with certain kinds of disabilities and for people taking the test in a language they are not fluent in.



Re-visit your purpose

The pilot is a good place to re-visit the purpose of your assessment and to check that through the planning, authoring and pilot stage, you have applied your purpose. Your goal should be that the assessment is valid, reliable and fair and that you have documented evidence to show this.

5. Best practice: Delivering the assessment

5.1 Test-taker communication and preparation

Here are some things to do before the actual assessment delivery takes place.

Provide opportunities for test-takers to practice with the testing user-interface

If a test-taker is unfamiliar with computers and/or with the user-interface for the exam, then their performance on the exam will be impacted by this lack of familiarity. They might reasonably claim that they needed time to understand the user interface and so any time limit is unfair, or they might even argue that they misunderstood the user interface and that their answers were unintended. To avoid such challenges, you should ensure that practice material is available using a similar user interface. It is common to provide practice questions with the same user interface as the real exam that test-takers can answer to familiarise themselves with the user interface.

Communicate appropriately around the use of the test-taker's personal information

Assessments usually collect personal data. In Europe, the General Data Protection Regulation imposes strong obligations regarding how personal data is handled and privacy laws also exist or being introduced in most other regions. Failing to comply with privacy law in your use of assessments can easily lead to legal challenge. For a clear description of the responsibilities of a data controller in Europe relating to assessments, please see Questionmark's white paper "Responsibilities of a Data Controller when Assessing Knowledge" available at **www.questionmark.com/learningresources**

Put in place an appeal process for test-takers

There are practical and legal reasons to have an appeals process. Practically, if there is an error or mistake in the exam process, your organisation will want to know about it and have a chance to resolve it before it gets escalated externally and impacts other test-takers. In some jurisdictions, there is value in having an appeals process, because the legal system encourages test-takers to exhaust such a process prior to going to court.

Case 6. Example of a poor appeals process

In summer 2018, an Irish student had the marks for her leaving certificate incorrectly calculated by the State Examinations Commission. If the marks had been correctly added up, she would have achieved the required achievement to be able to attend a university veterinary medicine course. But due to an error, the State Examinations Commission reported a lower score which did not allow the student to enter the course. This was the result of a simple, arithmetical error of not adding up the points correctly.

The student lodged an appeal under the appeals procedure, but this was very slow moving, and she did not get a response prior to the course cut-off date. She then raised an action in the Irish High Court to complain about the appeals process and request quicker resolution. The court ruled that the appeals process was "manifestly unfit for purpose" and directed the State Examinations Committee to quickly resolve the issue. It did so and the student was permitted to enter the course.

There were no complex issues of assessment validity or reliability here, just a simple mistake. Yet the student had to go to court with the associated high cost and stress in order to obtain relief. Additionally, the State Examinations Commission was the subject of unfavourable news coverage in the Irish and international press. This is an excellent example of why a sensible, properly functioning appeals process is in the interest of all stakeholders.

Present test-takers with an agreement or honour code

It is good practice to give test-takers an agreement setting out their rights and obligations in advance of an assessment. One benefit of this is that it can reduce cheating by reminding test takers of their responsibilities and of the possible consequences of behaviour that is contrary to the rules of the agreement. From a defensibility perspective, it may help with confirming that the test-taker understood and agreed to exam procedures at the time of taking the test.

Define consistent procedures for assessment delivery

A reliable assessment needs consistent procedures which should be documented and regularly reviewed. You should also ensure that proctors or any other assessment administrators are well trained. To quote ISO 10667-1:2011, it is important that ***"assessment administrators have the necessary competence based on verifiable experience, training, education or credentials and that, when administering an assessment to one or more individuals, assessment administrators follow the standardized procedures for the delivery of the assessment and document any deviations from those procedures."***[13]

Evaluate and mitigate security risks

An assessment that has been subject to widespread cheating is hard to defend. You should evaluate the risks to any exam and put in place measures to mitigate these risks. To learn more about assessment security, please see Questionmark's presentation "9 Risks to Test Security (and what to do about them)" [14]

13 ISO 10667□1:2011 section 5.4
14 This webinar is available at
http://pages.questionmark.com/WC-ENUS-WebinarRecording-9-Risks-Test-Security-201811_LP-Registration.html

5.2 Assessment delivery

Here are some key things to do during assessment delivery itself to reduce the chances of a legal challenge.

Authenticate test-taker

You do not want a test-taker to be able to claim that he/she never took the test and that someone else took it in his/her place without his/her knowledge. You should therefore put in place an effective authentication method for test-takers. For high stakes, proctored exams, it is usual practice to check at least one government issued photo ID. You should ensure that there is a consistent, reliable, documented process for this, and if personal data laws do not permit you to retain a copy of the ID that is checked, ensure that you have evidence that it was checked. If you are not using a government ID, ensure that your authentication method is documented and well-protected.

Consider proctoring

The prime motivation for proctoring is usually to reduce the risk of cheating. However, proctoring the assessment also provides useful evidence that the test-taker was present for the assessment and that the administration was conducted properly and consistently. Increasingly online proctoring is used, where a remote proctor observes the test-taker over a video link.

Present a clear user interface that displays the questions effectively

If a test-taker can argue that the user interface was confusing or ambiguous, or that they did not navigate through it as expected, this potentially means that the test could have been unfair in that it did not genuinely measure their skill or knowledge. It is obviously sensible to use a professional, high quality user interface.

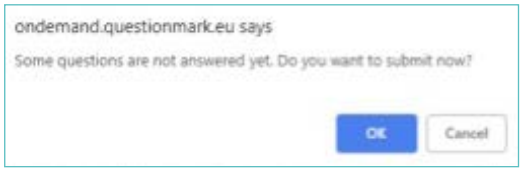
Ensure that the user interface allows test-takers to easily navigate between questions

A related issue is that if you are presenting a computerised exam as an alternative or equivalent to a paper one, then it is important that test-takers are allowed to navigate freely between questions, for example, to go backwards, forwards and flag a question to return to. This is possible in paper exams and so should be possible in computerised ones.

In some cases, you may be able to justify that the exam is "forwards only", i.e. questions must be answered in order and test-takers cannot return to previous questions - but this is less common and needs a rationale.

Ensure that the user interface requires test-takers to confirm before final submission

How would you deal with an appeal where the test-taker claims that they clicked the wrong button and submitted the test too early, thus they should be given another attempt at the test? To forestall such possible challenges and to give a good user interface to the test-taker, ask them to confirm before they submit their answers.



Save answers frequently in case of technical problems

Computers and connections fail from time to time. It is important to deliver exams in a way that can deal with such failures. An effective way to do this is to save answers to each question as test-takers answer them. This means that if there is a problem, all answers given until that point have been saved, and it is easy to resume the exam on another device after re-booting or when connection returns.

Ensure that all test-taker communication during exam delivery protected by TLS or equivalent

Suppose that a test-taker claims that they answered a question one way but their answer somehow was intercepted en route to the server and changed? You need to have evidence that this could not happen. In most environments, the easiest way to ensure this is to use TLS to set up a secure HTTPS connection, in the same way as electronic banking and other web systems use.

Most assessment software will use TLS, and you can check the quality of the communication using a test suite like at **www.ssllabs.com/ssltest**. Ensure that your environment gets a good score (A or A+).

Accessibility

You should ensure that your delivery software meets accessibility standards (often WCAG 2) and that assessments can be taken by those with disabilities, if the test specification allows it.

Capture evidence of test submission

It is important to use an assessment delivery mechanism that captures transactional details of when the test answers were made. For example, you would expect to capture the IP address and date/time of answers being submitted, and to have this information in a log file or database record that is protected against tampering. This will provide evidence in the event a test-taker claims he or she did not provide the answers that have been saved, or that he/she changed them later.

6. Best practice: Evaluating and reporting on the assessment

6.1 Reports and results management

Here are some measures in this area.

Ensure that results are tamper-proof

It is unusual but not unheard of for results to be tampered with after assessment completion. For example, USA Today has reported some US court cases where people were charged with altering grades by hacking into systems to change results after the event. A similar episode occurred in India recently. You should ensure that assessment results are reliably recorded, ideally in a read-only system that does not permit changes. If there is some additional system which collates and manages grades or scores, you should ensure it is tamper resistant.

If using human graders or observers, put in place robust measures to ensure reliability

If an observer is grading someone's performance, or if humans are grading test-taker work, it is important to put in place measures including rubrics, training and rater review to ensure consistency between raters. The reliability of an assessment can be significantly impacted by raters having different subjective views (inter-rater reliability), and you need to put in place, document and repeatedly carry out measures to reduce this.

Ensure a report is available which shows a full detail of an assessment attempt

If a participant disputes an assessment, the first thing you will want is to ensure that there is a report which contains:

- Each question in the test;
- The answers given by the participant;
- How the question was scored;
- A checkable calculation of how the test was scored and pass/fail determined;
- Audit trail information such as date/time, demographics, IP address and so on.

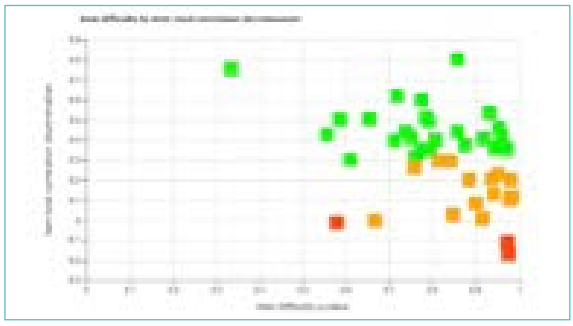
It should be possible to check or add up the score of a test manually, to demonstrate that no error has been made. Questionmark software provides a full report option within our “Coaching report” facility which includes all the above and more, and if you use another system, please ensure that it has all you need. It is also helpful if web logs or other transactional records are available to provide further evidence.

6.2 Evaluation of the assessment

Here are some of the things you should consider having in place for defensibility purpose. You may wish to do other evaluation for other purposes, e.g. cheater detection and analytics.

Run item analysis

Item analysis looks at the difficulty of questions (items) and how they correlate to test results. You should use item analysis to review the assessment periodically and identify poor items to either be improved or removed. A good report like the one produced by Questionmark on the right-hand side, will colour code questions and flag those that are potentially weak. Item analysis helps you to find questions which are:



- Too easy (and so do not meaningfully contribute to your assessment);
- Too hard (and potentially ambiguous or mis-coded);
- Confusing to strong test-takers or do not sufficiently distract weaker test-takers because of the available choices;
- Not correlative with test results, indicating that they may measure the wrong construct or be ambiguous

There are several useful resources on item analysis on the Questionmark's website[15]. It would be hard to defend an assessment where you do not run item analysis regularly.

15 See **www.questionmark.com/learningresources**

Evaluate the reliability statistics for the assessment

It is possible to statistically calculate reliability measures for the assessment, including the Chronbach's alpha, which is a measure of internal consistency of the test questions. Providing the test does measure a single construct, this is useful to calculate – both after a pilot and as a measure of reliability for an ongoing assessment. For a higher stakes test, psychometricians will calculate other measures.

Periodically review that the time limit for the exam remains fair

You should from time to time check that the time limit for the assessment remains fair (refer to section 4.3 for how to set a fair time limit).

Review question content to ensure it is still current

Put in place a regular review mechanism for questions to check if they need retiring, either because they become out of date, or because they become exposed (e.g. communicated on the Internet) and so get too well known.

Evaluation in relation to the purpose of the assessment

Lastly, ISO 10667 requires (and it is good practice for all to have) a process of evaluating the assessment including reviewing errors or problems, ensuring that good practice is being followed (e.g. your procedures are being implemented), that the scoring criteria are still relevant and that the equity or fairness of all relevant sub-groups is maintained.

Ultimately this evaluation and the whole purpose of your programme is to ensure that the assessment remains valid, that it is to say the scores are useful for the purpose that you plan to use them.

We hope that you have found this best practice guide useful to help you create and administer defensible assessments that contribute to legal certainty. The authors would welcome all comments on how to improve this guide, any input is welcome to John Kleeman at **john@questionmark.com**.

Appendix – Further information

Here are some further information resources that would be helpful to people seeking good practice in defensibility and legal certainty in the area of assessments.

Books, white papers and articles

Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training, by Sharon A. Shrock and William C. Coscarelli, 3rd edition, 2007, Pfieffer, an excellent book on test development.

Documenting Psychometric Evidence: A Goldilocks Conundrum, a 2019 white paper published by the Institute for Credentialing Excellence (ICE).

Performance Based Certification. How to design a Valid, Defensible, Cost Effective Program by Judith Hale, 2011, Pfeiffer, an expert book.

Rechtliche Aspekte von E-Assessments an Hochschulen (Legal aspects of E-Assessments in Universities) available at **http://duepublico.uni-duisburg-essen.de/servlets/DocumentServlet?id=42871**, a 2016 analysis of legal aspects of the use of electronic assessments in German universities (written in German only).

Defensible Assessments: What You Need to Know, a 2007 white paper by Questionmark that gives useful context around defensible assessments, available at **https://www.questionmark.com/wc/WP-ENUS-Defensible-Assessments**

Standards and guidelines

ISO 10667:2011, Assessment service delivery - Procedures and methods to assess people in work and organisational settings.

ISO 23988:2007, Information technology - A code of practice for the use of information technology (IT) in the delivery of assessments.

International Test Commission guidelines at **https://www.intestcom.org/page/5**

NCCA Standards available from the Institute for Credentialing Excellence at **www.credentialingexcellence.org**

Standards for Educational and Psychological Testing developed jointly by the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (see **www.apa.org/science/programs/testing/standards.aspx**).

Principles for the Validation and Use of Personnel Selection Procedures, Fifth Edition, 2018, available at **https://www.apa.org/ed/accreditation/about/policies/personnel-selection-procedures.pdf**

Readers will also find a lot of useful information on writing good quality tests and exams on the Questionmark blog at https://blog.questionmark.com and on Questionmark's website at **https://www.questionmark.com/learningresources**

www.questionmark.com
sales@questionmark.com

USA
35 Nutmeg Drive, Suite 330
Trumbull CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
Moor Place, 1 Fore Street
London EC2Y 9DT
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+ 49 (0) 89 220 61272

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

www.questionmark.com

sales@questionmark.com

USA
35 Nutmeg Drive, Suite 330
Trumbull CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
Moor Place, 1 Fore Street
London EC2Y 9DT
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+ 49 (0) 89 220 61272

WHITE PAPER

Questionmark provides a secure enterprise-grade assessment platform and professional services to leading organizations around the world, delivered with care and unequalled expertise. Its full-service online assessment tool and professional services help customers to improve their performance and meet their compliance requirements. Questionmark enables organizations to unlock their potential by delivering assessments which are valid, reliable, fair and defensible.

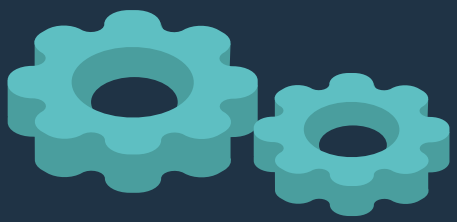
Questionmark offers secure powerful integration with other LMS, LRS and proctoring services making it easy to bring everything together in one place. Questionmark's cloud-based assessment management platform offer rapid deployment, scalability for high- volume test delivery, 24/7 support, and the peace-of-mind of secure, audited U.S., Australian and European-based data centers.

Questionmark has the experience to ensure that its customers get results they can rely on. It has helped its customers deliver more than 95m unique assessments and, since starting, has been trusted by more than 2,500 customers worldwide.

The business has a wide range of expertise across industry sectors, government and academia. These include, but are not confined to, financial services, technology, pharmaceuticals, utilities, retail, public sector and government, awarding bodies and higher education.

Questionmark has achieved authorization from the Federal Risk and Authorization Management Program (FedRAMP). The FedRAMP Authorization means Questionmark is approved to deliver cloud-based assessments for the US government and military organizations.

The business supports the full range of roles within customers' organizations to deliver valid, reliable, fair, and defensible assessments. This includes senior managers and departmental heads, technical assessment teams, and consultants and intermediaries. Questionmark also supports in-house functions such as IT, data, legal and procurement teams.



www.questionmark.com
sales@questionmark.com

USA
35 Nutmeg Drive, Suite 330
Trumbull CT 06611
+1 (203) 425 2400
(800) 863 3950

UK
Moor Place, 1 Fore Street
London EC2Y 9DT
+44 (0) 20 7263 7575
0800 731 5895

Germany
Hopfenstr. 8,
80335, Munich
+ 49 (0) 89 220 61272