The Legal Defensibility of Assessments: What You Need to Know

This paper explores legal defensibility in the area of assessment, describing how Questionmark Perception can be used to help ensure the legal defensibility of an assessment program, and help an organization meet the needs of international standards such as ISO/IEC 17024. This paper is designed to help readers determine what legal defensibility means, provide readers with direction in terms of the standards and best practices available, describe how to ensure and evaluate legal defensibility, and discuss specific examples of how Questionmark Perception software helps with legal defensibility.

Author: Greg Pope

With assistance from: Eric Shepherd
John Kleeman
Brian McNamara
Joan Phaup

Table of contents

Executive Summary

Legal defensibility has to do with the ability of a testing entity (e.g., certification organization) to withstand legal challenges. These legal challenges may come from individuals or groups who claim that the organization 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, legal defensibility has to do with the question: “Are the assessment results and, more generally, the testing program defensible in a court of law?”

Even if your organization is not directly concerned by legal threats to your testing program, following good practice in this area is likely to make your testing program fairer, more valid, and more useful.

- ► This document references generally accepted guidelines in the area of psychometrics (educational and psychological measurement), including international standards, and these are referenced in this document.

According to legal experts and recorded legal cases, the four main areas that tests are typically legally challenged on are reliability, validity, fairness, and cut scores. Each of these areas requires consideration when developing, administering, and reporting (conducting research) on assessment results. In medium and certainly in high-stakes testing scenarios, focusing on these four major areas will help ensure accuracy and fairness in testing.

Cheating on tests and breaches of security are areas of concern that throw into question the validity of assessment results. If assessment results cannot be trusted to be an accurate reflection of a participant’s “true” ability or performance level, then the validity of results is questionable and the legal defensibility of the assessment program may be in jeopardy. Security is important to ensuring legal defensibility and can be enhanced by Questionmark products.

Legal defensibility requires vigilance both in terms of process and information use. Assessment organizations should assemble and maintain a portfolio of legal defensibility evidence in order to promote best practices and document processes and procedures. Using Questionmark products to help ensure legal defensibility will benefit both the assessment organization administering tests and the participants taking those tests.

1. What is legal defensibility?

Legal defensibility, in the context of assessment, has to do with the ability of a testing entity (e.g., certification organization) to withstand legal challenges. These legal challenges may come from individuals or groups who claim that the organization 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, legal defensibility has to do with the question: "Are the assessment results and, more generally, the testing program defensible in a court of law?"

The consequences of not having legally defensible practices—both to participants and to the testing organization—are great. In high-stakes assessment, participants who are inappropriately certified can jeopardize the stakeholders they serve (e.g., doctors who obtain their certifications without being truly qualified can put their patients at risk). Participants, who deserve to be certified but are turned down, may have grounds for legal action against the certifying body. It is also a loss to the participant's profession (e.g., medicine) if qualified people are not being certified. In terms of certifications, the decision to certify or not certify is generally partially based on a cut score (e.g., 50% correct on an assessment). A difference of even one percentage point on an assessment (e.g., 50% versus 51%) can result in some qualified people not being certified or unqualified people being certified inappropriately.

In low- and medium-stakes assessment, the risks may appear to be less, but the effects of not following legally defensible practices, both on participants and organizations, can also be felt. The reputation of the organization can be tarnished when participants lose faith in the quality of the assessments they are taking. More appeals will occur if participants feel that the tests are not measuring what they are designed to measure or if the quality of the questions is called into doubt. Course evaluation surveys will reflect dissatisfaction among participants. Following best practices in the assessment process takes time, but it will pay dividends in the form of higher quality assessments. Following best practices and standards, regardless of the purpose of the assessment program, will have benefits for developers, stakeholders, and participants.

Legal systems vary from country to country. While this white paper focuses mostly on issues that have arisen within the U.S. legal system, the principles are likely to be useful worldwide. Although this document addresses issues with regard to legal defensibility, this paper is meant for general education purposes only; in no way should it be considered a substitute for legal advice from qualified legal counsel.

Some examples of legal cases, in the area of assessment legal defensibility, are outlined below.

- United States versus Delaware (2004 WL, 609331, D. Del. March 22, 2004) showed that the employer must justify the process of setting a cut score.
- International Brothers of Electrical Workers versus Mississippi Power (442 F.3d, 313 - 5th Cir. 2006) held that the employer was justified in raising a pass score

as the employer showed that by doing so the employees who were passed demonstrated above-average performance on the job.

- Tolleson versus Educational Testing Service (832 F. Supp. 158, D.S.C. 1992) determined that it was essential to follow best practices and due process when investigating incidences of cheating.
- Mexican-American Educators versus California (84 FEP Cases 474, 9th Cir. 2000) related to ensuring equal test performance on the California State teachers credentialing exam.
- Pratt versus California State Board of Pharmacy (No. Civ. 05-0345 DFLPAN, E.D. Calif. 2006) rejected a challenge to a passing score based on the modified Angoff standard setting procedure used and the standard error of measurement.

Note: Above examples were referenced from Abram, T. G. (2007), *Avoiding Legal Challenges To Testing Programs: A Few Ounces of Prevention,* Association of Test Publishers Annual Meeting, Palm Springs, CA.

Adherence to best practice guidelines and standards in the area of psychometrics (educational and psychological measurement) is essential for an assessment process to be considered legally defensible. See Chapter 2 for more information on these standards and best practices.

In the event of legal challenge, the assessment organization would need to provide evidence that professional guidelines/standards have been adhered to (from planning and content development to administration, scoring, and reporting).

According to legal experts and recorded legal cases, the four main areas that tests are legally challenged on are: reliability, validity, fairness, and cut scores. These four areas (discussed in more detail in Chapter 3) are summarized below.

- *Reliability* has to do with how consistently the test measures a construct (a construct is the latent variable that is being assessed, such as mathematical ability or mechanical aptitude).
- *Validity* generally refers to whether the test is measuring what it is supposed to measure.
- *Fairness* refers to the test only measuring the construct(s) it was designed to measure with no unfair advantage for any given demographic group or subpopulation.
- *Cut scores* are the “pass/fail” benchmarks generally used to determine whether participants have demonstrated an appropriate level of knowledge or skill on a test.

As the testing industry moves more toward computerized authoring, scheduling, delivery, and reporting, tools such as Questionmark Perception can help ensure legal defensibility. The purpose of this paper is to present best practices in the area of assessment and describe how Questionmark Perception can help toward this end.

2. Guidelines, standards, and best practice resources

For an assessment process to be considered legally defensible, best practice guidelines and standards in the area of psychometrics (educational and psychological measurement) must be followed. Many best practice guidelines and standards exist and are produced by various organizations all over the world. This chapter provides background information to set the context for specific discussion later in this resource.

2.1. Standards for Educational and Psychological Testing (1999)

www.apa.org, www.aera.net, www.ncme.org

An important publication, *Standards for Educational and Psychological Testing* (1999), was developed jointly by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The publication is intended for professionals as well as the layperson in the area of assessment. It focuses on test development in the areas of education, psychology, and employment. Although produced in the United States, the publication recommends practices that are applicable internationally.

Standards for Educational and Psychological Testing (1999) presents best practices for a broad audience geared towards organizations developing educational (e.g., state or provincial level achievement) and psychological (e.g., personality) tests at various levels of stakes (low to high) but that have applicability to any area of assessment (e.g., licensure and certification).

An outline of the content of this publication is provided below.

Part I: Test Construction, Evaluation, and Documentation

- Validity
- Reliability and Errors of Measurement
- Test Development and Revision
- Scales, Norms, and Score Comparability
- Test Administration, Scoring, and Reporting
- Supporting Documentation for Tests

Part II: Fairness in Testing

- Fairness in Testing and Test Use
- The Rights and Responsibilities of Test Takers
- Testing Individuals of Diverse Linguistic Backgrounds
- Testing Individuals with Disabilities

Part III: Testing Applications

- The Responsibilities of Test Users
- Psychological Testing and Assessment
- Educational Testing and Assessment

- Testing in Employment and Credentialing
- Testing in Program Evaluation and Public Policy

2.2. Conformity assessment – general requirements for bodies operating certification of persons 17024 (2003)

www.iso.org

Developed as an international standard and made available in the United States by the American National Standard Institute (ANSI), this document provides guidelines and standards in the area of certification.

The 17024 document is structured into succinct numbered sections that allow an organization to easily conduct an audit to determine whether it is conforming to the guidelines. For example, an organization could create a checklist in which each section of the 17024 standards document is compared with how the organization currently meets each of the requirements. This provides an audit of an organization's current status in relation to the standards and a starting place for developing a roadmap for improvements, if required.

An outline of the content of this publication is provided below.

- Section 1: Scope
- Section 2: Normative references
- Section 3: Terms and definitions
- Section 4: Requirements for certification bodies
    - 4.1: Certification body
    - 4.2: Organizational structure
    - 4.3: Development and maintenance of a certification scheme
    - 4.4: Management system
    - 4.5: Subcontracting
    - 4.6: Records
    - 4.7: Confidentiality
    - 4.8: Security
- Section 5: Requirements for persons employed or contracted by a certification body
    - 5.1: General
    - 5.2: Requirements for examiners
- Section 6: Certification process
    - 6.1: Application
    - 6.2: Evaluation
    - 6.3: Decision on certification
    - 6.4: Surveillance
    - 6.5: Recertification
    - 6.6: Use of certificates and logos/marks

2.3. Guidelines for Computer-Based Testing (2000)

www.testpublishers.org/

The Guidelines for Computer-Based Testing (2000) is a publication produced by the Association of Test Publishers (ATP), of which Questionmark is an active member.

The audiences for the Guidelines for Computer-Based Testing (2000) are broad and include:

i. Test development organizations
ii. Test publishers, test administrators, and test delivery organizations
iii. Licensure and certification boards

An outline of the content of this publication is provided below.

Part 1: Background & Explanations

- Chapter 1: Introduction
  - Audiences for the Guidelines
  - Computer-based Testing
  - Scope of the Guidelines
  - Overview
- Chapter 2: Validity and Test Design
  - Planning the Test
  - Test Specifications
  - Evaluation Plan
- Chapter 3: Test Development and Analysis
  - Item Banking
  - Test Assembly
  - Test Fairness
  - Item and Test Analysis
  - Tracking and Monitoring
- Chapter 4: Test Administration
  - Guidelines for Test Administration

Part 2: Computer-Based Testing Guidelines

- Chapter 1: Planning and Design
- Chapter 2: Test Development
- Chapter 3: Test Administration
- Chapter 4: Scoring and Score Reporting
- Chapter 5: Psychometric Analysis
- Chapter 6: Stakeholder Communications
- Chapter 7: Security

2.4. Other guidelines and resources

- *International Test Commission (ITC) Projects* http://www.intestcom.org/itc_projects.htm These projects are translated into several languages and provide best practices in several areas and are intended for any organization involved in low-, medium-, and high-stakes assessment.
    - ITC Guidelines on Adapting Tests
    - ITC Guidelines on Test Use
    - ITC International Guidelines on Computer-Based and Internet-Delivered Testing
- *Principles for Fair Student Assessment Practices for Education in Canada* http://www.education.ualberta.ca/educ/psych/crame/files/eng_prin.pdf This document is intended for students, parents, teachers, and organizations involved in education-based assessment.
- *Gender and Fair Assessment* (Willingham and Cole, 1997). This book provides a great deal of information on the theory and practice of gender in the area of assessment and is intended for anyone interested in this topic.
- *Code of Fair Testing Practices in Education* http://www.apa.org/science/fairtestcode.html This publication applies broadly to testing in education (admissions, educational assessment, educational diagnosis, and student placement) for all modes of presentation (computer, paper-and-pencil, etc.).
- *NCTA Professional Standards and Guidelines for Post-Secondary Test Centers* http://www.ncta-testing.org/resources/standards/standards.php This resource is intended for organizations that conduct testing at the post-secondary level and require information on test center best practices.
- *Information technology—A code of practice for the use of information technology (IT) in the delivery of assessments (ISO/IEC 23988:2007)* www.iso.org Designed for groups or individuals who conduct assessments using IT platforms and are interested in knowing best practices.
- *Standards for Teacher Competence in Educational Assessment of Students* (1990) http://www.unl.edu/buros/bimm/html/article3.html This document is intended for use by teachers and those who develop professional development experiences for teachers.
- *Guidelines for Computerized Adaptive Test Development and Use in Education* (1995) www.acenet.edu This resource is designed for those involved in the development of computerized adaptive testing (CAT) programs.

3. Some areas of legal challenge

According to legal experts and recorded legal cases, the four main areas that tests are legally challenged on are: reliability, validity, fairness, and cut scores. These areas will be focused on in this chapter.

A good general book for further reading on constructing assessments is *Criterion-referenced Test Development: Technical and Legal Guidelines for Corporate Training,* by Schrock & Coscarelli (2000).

3.1. Reliability

"Reliability" is used in everyday language: "My car runs reliably" means that it starts every time, it doesn't stall when driving, and so on. In the context of assessment, the term reliability refers to test score reliability in how *consistently* the assessment measures something (e.g., how well an assessment measures a "construct," such as math ability).

In general terms, an assessment is a measurement instrument composed of many individual measurements (questions/items). All measurement instruments (including assessment questions) have error in their estimates, so the traditional view of test score reliability states that an individual's observed assessment score equals his or her theoretical (un-measurable) "true" score plus error. Put another way, if an assessment theoretically had no error and a participant's memory of taking a test could be erased before taking the test again, then every time the participant took the assessment, he or she should obtain the same score. The test score reliability in this theoretical situation would be perfect (i.e., 1.0) because participants who take the assessment always get the same score (100% reliable measurement). In the real world, this is not the case; questions do not perfectly measure how much participants know and can do and therefore test scores are never perfectly reliable.

There are many approaches for measuring reliability. These include:

- Internal consistency: How well do the questions on the test "hang together"?
- Split-half (split-forms): Comparison between of two forms (splits) of the same test (first 25 items versus last 25).
- Test-retest: Comparison between multiple administrations of the same test.
- Inter-rater reliability: Comparison between two or more raters (markers) who rate the same thing (e.g., provide essay scores).

The most common approach to measuring test score reliability is the internal consistency reliability approach, of which the Cronbach's Alpha method is quite common. Cronbach's Alpha was designed for use with questions that are scored as right/wrong (e.g., true/false) as well as partial credit (e.g., essay questions, 0–5). The

Cronbach's Alpha coefficient is the same for dichotomous questions (right/wrong) as the KR-20 reliability coefficient (another internal consistency reliability measure).

The typical range of Cronbach's Alpha is between 0 and +1, although values outside this range (e.g., negative values) are possible in some situations. Reliability values closer to +1 indicate higher reliability and values above 0.90 are generally considered acceptable for high-stakes assessment (where an assessment covers a single topic).

Factors that influence the internal consistency reliability of test scores are:

- Test length: Generally more items on the assessment result in higher reliability as more measurements of the construct are obtained.
- Item discrimination (e.g., item-total correlations): Items with higher discrimination (fidelity) provide more measurement value (depends on what is being measured and who is being measured: for example, it may not be the item that is flawed, maybe the class was not taught the material properly).
- Item difficulty: Items that are extremely hard or extremely easy affect discrimination and therefore reliability. If a large number of participants do not have time to finish the test, this affects item difficulty.
- Construct being measured: If all questions are measuring the same construct (e.g., from the same topic) reliability will be increased.
- How many participants took the test: With very small numbers of participants the reliability value will be less stable.

Who should be investigating/using reliability analysis methods?

- Must: High-stakes assessment organizations (certification, accreditation, entrance exams, etc.)
- Should: Medium-stakes (college tests, skills assessment)
- Could or not applicable: Low-stakes (self assessment, quizzes, tutorials)

The Item Analysis Report and Test Analysis Report available within Perception generate Point-Biserial/Pearson item-total correlations for questions and Cronbach's Alpha for an assessment. These reports and suggestions for how to use them to address legal defensibility are discussed in more detail in Chapter 4.

3.2. Validity

Validity deals with whether the test measures what it is supposed to measure. The process of validation deals with gathering evidence about the following:

- Whether the test is measuring what it should be measuring
- How well the processes followed to create the test and test questions are subscribing to the generally accepted best practice guidelines and standards
- Whether experts have opportunities to review and rubber stamp the processes

- The ability of the test results to predict the intended outcomes

Traditionally, validity refers to three areas:

1. Criterion-related (predictive/concurrent): Are scores related to outcome measures (e.g., do secondary school exam scores predict post-secondary performance)?
2. Content-related: Is the test content appropriate for what is being measured (e.g., is content coverage representative of the curriculum)?
3. Construct-related: Is the test measuring the correct trait, attribute, ability, or construct (e.g., is this math ability test measuring math ability)?

3.2.1. Criterion-related validity

Criterion-related validity relates to the degree to which variables (e.g., participant assessment scores) predict an outcome (e.g., on the job performance). Therefore, criterion-related validity involves comparing scores with a criterion. Criterion-related validity is sometimes divided into four subtypes: predictive, concurrent, convergent, and discriminate.

Predictive validity refers to whether the assessment scores predict something they ought to predict. For example, college entrance exam scores should predict performance in post-secondary education. Another example may have to do with whether mechanical aptitude test scores predict on the job performance for automotive mechanics.

Concurrent validity deals with whether the assessment can differentiate between different groups of participants. For example, professors in mathematics should perform better than junior high school students on a high school mathematics achievement test.

Convergent validity refers to the degree to which an assessment's scores are similar to other assessment scores that measure similar things. For example, a mechanical aptitude test developed by one group should yield results similar to those of a mechanical aptitude test developed by another group (so long as the two tests are measuring similar sub-constructs).

Discriminate validity is similar to convergent validity but, instead of obtaining evidence that assessment scores are similar between assessments measuring similar constructs, discriminate validity examines whether assessments measuring unrelated constructs are indeed not related. For example, an assessment of mechanical aptitude should have a low correlation (i.e., little or no relationship) with an assessment of fashion aptitude. A low correlation between the two provides evidence of discriminate validity.

3.2.2. Content-related validity

Content-related validity refers to whether the content of an assessment is appropriate to what is being measured (i.e., is a mathematics test composed of mathematics questions related to the curriculum). Subject matter experts (SME) and other experts generally review tests to ensure that the material—the content coverage of questions—is appropriate for what is being tested. If a blueprint or plan for the assessment was created and vetted, and test designers adhered to the blueprint or plan, content validity should be acceptable.

3.2.3. Construct-related validity

Construct-related validity examines whether the assessment measures the construct that it was designed to measure. This is generally determined by investigating the statistical characteristics of the assessment such as the reliability, dimensionality, and so on. Dimensionality has to do with whether an assessment is measuring the constructs or topic(s) that is was designed to measure.

A crucial aspect of validity, which is also tied into reliability, has to do with the quality of questions being developed. Good quality questions are unambiguous, clearly written, and focused on the construct. They avoid some of the traps that are possible when writing multiple choice distracters, giving them a higher likelihood of being deemed to be valid and a greater chance of contributing positively to the reliability of the assessment.

At times, people refer to “face validity.” This generally refers to whether “on its face” the assessment is measuring the construct it was designed to measure (does a test designed to measure mechanical aptitude “look like” it is measuring this construct?). Although face validity may be important for the perception of credibility and “buy-in” from various groups, it is generally not considered a robust or formal method of determining validity in a legal defensibility context. With this said, however, if an assessment does not look like it is measuring what it is supposed to be measuring, an organization could expect many appeals from participants and questions from stakeholders.

A common question is whether reliability and validity are related and if so how. The brief answer to this question is that an assessment must be reliable in order to be valid and that an assessment can be highly reliable but not valid. Here is an example of how an assessment can be highly reliable but not valid:

> As long as an assessment measures the same thing consistently, its reliability will be high. Most people can remember their name reliably, most people can remember their height reliably, and so on, so a test composed of 100 questions that all asked these types of questions would be highly reliable because participants would answer the questions consistently. If the test was supposed to measure mechanical aptitude, but was composed of questions regarding height,

weight, and so on, the reliability of the test would be high but the validity of the test would be low.

3.3. Fairness (and bias)

Bias, in the context of assessment legal defensibility, refers to whether the test performs fairly for different groups or demographics. Fairness refers to how the results are used.

Generally, a test is regarded as biased if it performs differently for different groups or demographics that have the same level of ability (e.g., do the top participants in each demographic group perform significantly differently, do the middle participants in each demographic group perform significantly differently, etc.). Many factors can contribute to bias in test scores, including:

- How the test was administered
- The translation of the test questions
- Content of the questions (e.g., graphs that are not familiar, understandable, or even offensive to certain groups or demographics)

Fairness refers to how the results (biased or unbiased) are used. For example, test results for a test that has been found to be free of bias may be used unfairly by applying different standards or criteria (e.g., cut scores) to different demographic groups. Fairness should be addressed by following best practice guidelines in the area of assessment and conducting independent reviews of processes to ensure that the decisions being made are as fair as possible.

Bias and fairness should be examined in conjunction as one may hold clues to the other. To cite a hypothetical situation: An analysis by the human resources department found that people already in a job at an organization performed very similarly on the job regardless of demographic group. Yet test score differences (bias) were found between demographic groups for new applicants who had to take a new test (composed of all English questions) as part of the application process. One reason for this discrepancy between on the job performance (fairness) and test scores (bias) could be language differences. Reading level and comprehension of the language of the test can be a major contributor to group or demographic differences on test results, resulting in a fairness issue because the test result differences do not reflect the applied on-the-job performance. One step to remedy this situation could be to translate the test into other languages that correspond to the most common first languages of the people taking the tests. This would make the results more usable by the organization, address the bias in the test, and improve the fairness of the process.

3.4. Cut scores

Cut scores are the benchmarks (e.g., “pass/fail”) generally used to determine whether participants have demonstrated an appropriate level of knowledge or skill on a test. For

example, a “pass score” of 50% on a test would mean that participants would be required to achieve a score of at least 50% in order to pass the test. Cut scores do not only refer to pass/fail benchmarks: organizations may have several cut scores within an assessment that differentiate between “Advanced,” “Acceptable,” and “Failed” levels.

Cut scores are very common in high and medium-stakes assessments and well established processes for setting these cut scores and maintaining them across administrations are available (i.e., standard setting procedures and statistical equating techniques). Generally, one would first build/develop the assessment with the cut score in mind. This would entail selecting questions that represent the topics being covered, considering the distribution of difficulty of the questions, selecting more questions in the cut score range to maximize the “measurement information” near the cut score, etc. Once a test form is built or a question repository is established, this test form or repository would generally undergo formal standard setting procedures to set or confirm the cut score(s). There are many standard setting methods used to set performance levels and cut scores on tests, generally split into two types: a) test/question-centered approaches and b) participant-centered approaches. A few of the most popular methods, with very brief descriptions of each, are provided below (for more information on standard setting procedures and methods, see Cizek, 2001).

- *Modified Angoff method* (test/question-centered): Subject matter experts (SMEs) are generally briefed on the Angoff method and allowed to take the test with the performance levels in mind. SMEs are then asked to provide estimates for each question of the proportion of borderline or “minimally acceptable” participants that they would expect to get the question correct. Several rounds are generally conducted with SMEs allowed to modify their estimates given different types of information (e.g., actual participant performance information on each question, other SME estimates, etc.). The final determination of the cut score is then made (e.g., by averaging estimates). This method is generally used with multiple-choice questions.
- *Nedelsky* method (test/question-centered): SMEs make decisions on a question-by-question basis regarding which of the question distracters they feel borderline participants would be able to eliminate as incorrect. This method is generally used with multiple-choice questions only.
- *Bookmark method* (test/question-centered): Questions are ordered by difficulty (e.g., item response theory b-parameters) from easiest to hardest. SMEs make “bookmark” determinations of where performance levels (e.g., cut scores) should be. (As the test gets harder, where would a participant on the boundary of the performance level not be able to get any more questions correct?) This method can be used with virtually any question type (e.g., multiple-choice, multiple-response, essay, etc.).
- *Borderline groups method* (participant-centered): A description is prepared for each performance category. SMEs are asked to submit a list of participants whose performance on the test should be close to the performance standard (borderline). The test is administered to these borderline groups and the median

test score is used as the cut score. This method can be used with virtually any question type (e.g., multiple-choice, multiple response, essay, etc.).

- ► *Contrasting groups method* (participant-centered): SMEs are asked to categorize the participants in their classes according to the performance category descriptions. The test is administered to all of the categorized participants and the test score distributions for each of the categorized groups are compared. Where the distributions of the contrasting groups intersect is where the cut score is located. This method can be used with virtually any question type (e.g., multiple-choice, multiple response, essay, etc.).

Questionmark Perception can help set cut scores to facilitate legal defensibility in a number of ways:

- → The Item Analysis Report can be used to obtain question difficulty information (e.g., p-values) for all participants or for certain groups of participants. This information can be used in a standard setting process to provide feedback to SMEs during test question centered standard setting sessions.
- → The Angoff field within Perception can be used to capture the question difficulty estimates when the Angoff standard setting method is being used.
- → When using the Bookmark method, the questions from the actual assessment can be ordered in a separate "Bookmark Assessment" created specifically for this purpose. The question level difficulty information from the Item Analysis Report can be used to inform the ordering of the questions from most difficult to least difficult.

3.5. Other issues

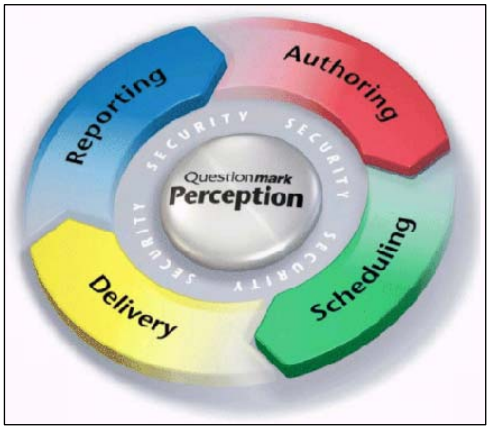
Adverse impact is an important and frequently-discussed issue related to high-stakes assessment. Adverse impact refers to a selection processes/criteria that select individuals in a demographic group at a higher rate than individuals in other protected demographic groups. The rate that is generally quoted is the "4/5 rule" or "80% rule," which states that if selection of a protected group occurs at a rate of less than 80% of the rate for the group with the highest selection rate, this may constitute evidence of adverse impact.

More information on adverse impact can be found on the U.S. Equal Employment Opportunity Commission web site (http://www.eeoc.gov/).

4. Questionmark products and legal defensibility

This chapter outlines how Questionmark products can help ensure legal defensibility in the four areas of the assessment process:

i. *Authoring* – The process of composing a bank of questions and then selecting appropriate ones for assessments that are given to participants
ii. *Scheduling* – The process of specifying which participants can take which assessments and when they can take them
iii. *Delivery* – The process by which participants receive their assessments
iv. *Reporting* – The process of reporting on and analyzing results



4.1. Authoring

The assessment process begins with the creation of high-quality test questions. Questionmark Authoring Manager offers 22 different types of questions (http://www.questionmark.com/us/perception/authoring_windows_qm_qtypes.aspx). This level of flexibility enables use of the most appropriate question types to effectively measure the constructs that are being tested. For example, a multiple-choice question may not be the best question type choice when the construct attempting to be measured is "writing skills." An essay question or short answer question type would be more suitable for measuring this construct. As such, one must choose the question type that best allows for the measurement of the construct being tested.

Scoring of questions is an equally important consideration when selecting the appropriate question type. Perception allows for a great deal of flexibility in the scoring of questions to ensure that questions are being scored in a psychometrically appropriate and fair manner.

For more information on best practices related to question types, topics, and assessments, see the guides included with Perception Authoring Manager (Help > Best Practice and Guidance).

For more information regarding Perception authoring software see: http://www.questionmark.com/perception/help/v4/manuals/

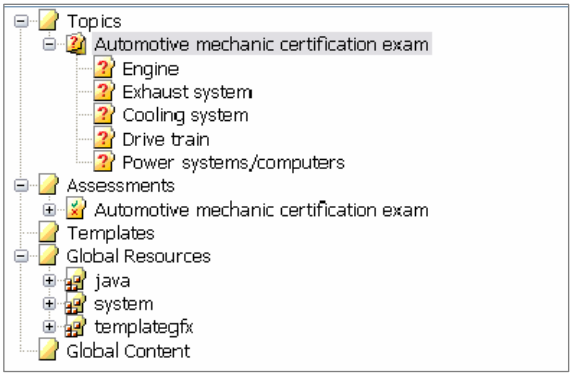
An important aspect of question development is the ability for others involved in the assessment program to review and edit the questions before they are administered to participants. This is an important step in ensuring the content-related validity of the

assessment. For example, subject matter experts should be reviewing each question to ensure that the content is focused to the specifications (e.g., blueprint) and program of study. Some people who may be involved in the review process are:

- Other subject matter experts
- Professional editors
- Psychometricians
- Copyright specialists
- Graphic design artists

Items and assessments in Perception are stored in repositories that can be accessed remotely and securely via the Internet and intranets.

- This virtual access and review process increases the legal defensibility of the testing program by ensuring that the most appropriate and qualified people review the questions, regardless of their geographical location.
- Questions within a repository can be appropriately organized into a topic structure. This is important for content-related validity as questions should be categorized according to their measurements of the constructs of interest. This allows for straightforward review of questions by SMEs and others as well as for efficient test form development.

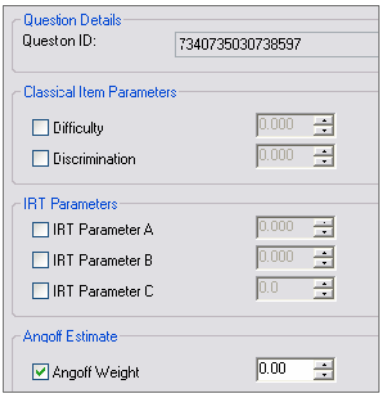


Security is an important issue in terms of legal defensibility. All best practice guidelines and standards documents discuss security because security has an impact on the validity of assessment results and fairness to participants. Perception provides extensive, customizable security capabilities in its authoring environment.

- Access to Questionmark repositories and security rights are role-based, helping to ensure the legal defensibility of the testing program from a security point of view. It is possible to set access permissions by role or profile that limit an author's or reviewer's access to the various functions within Perception.
- Administrators may limit which topic (item) folders and assessment folders authors may access, as well as what the author may do in those folders (e.g. view the contents only, view and edit the contents only, etc.)

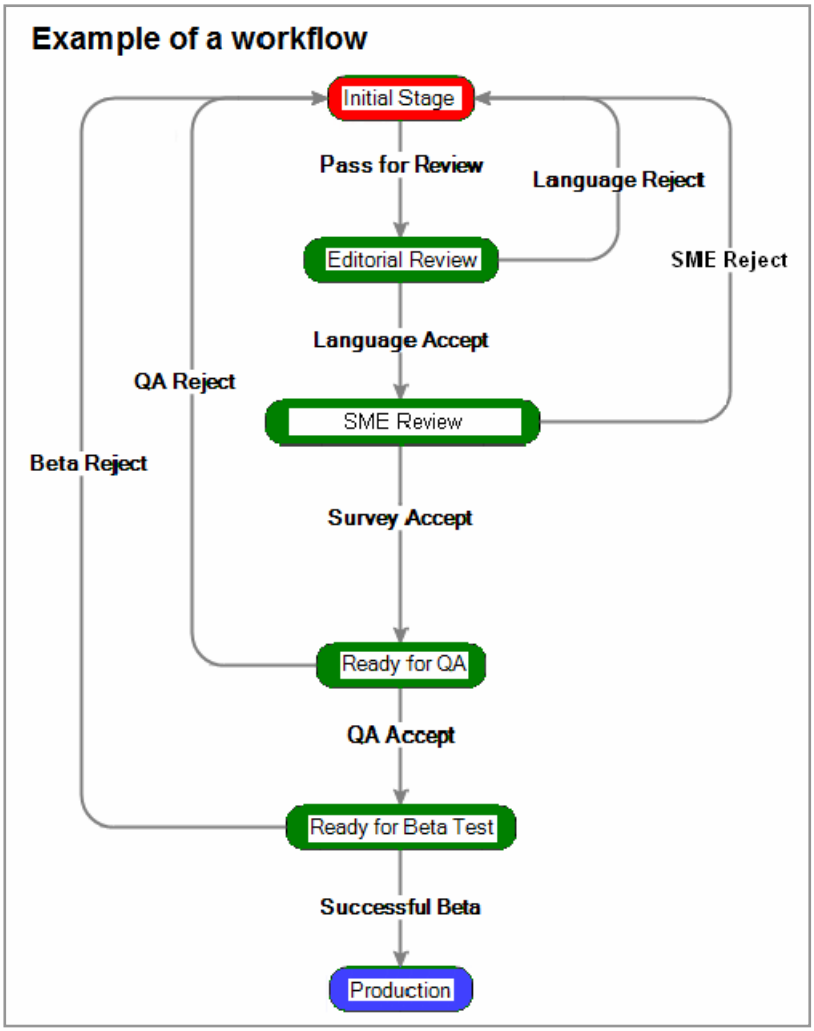
The translation of questions into multiple languages is necessary for many high-stakes tests in order to fulfill the legal defensibility guidelines for fair and equitable access. Questionmark Perception itself is available in multiple language formats and allows authors to develop questions in virtually any language using Unicode character sets. See the Questionmark white paper on translating questions: http://www.questionmark.com/catalog/help/resources/Best_Practice_Guide_for_Translating.pdf

When standard setting procedures are used, it is useful from a record keeping perspective to store information regarding the aggregate (e.g., mean or median) Angoff estimates (weights) for questions that SMEs have provided. A field is available in the question authoring process that can capture this information for each question. In addition, IRT parameter estimates can be entered for each question in order to record the information for record keeping and future analysis.



Using workflows provided within Authoring Manager helps to ensure legal defensibility of the assessment process by implementing structured stages and actions in a shared repository to guide the question development, review, administration, and analysis processes.

Using the question delivery status of "Normal," "Experimental," "Beta," "Incomplete," or "Retired" will help ensure legal defensibility by ensuring that questions are properly tracked and categorized throughout the assessment process.



For more information on workflows see: http://www.questionmark.com/perception/help/v4/knowledge_base/workflow/work401.aspx).

In order to ensure that the assessment is not performing differentially for different demographic groups, Perception assessments can be designed to capture participant demographic information and save this data to special fields in the assessment answer database. This information can be used to produce reports to help ensure equivalence of results.

4.2. Scheduling

Scheduling assessments to ensure that the right people are taking the tests, and that sound security measures are in place, is vital to legal defensibility. Administrators have many options when it comes to scheduling assessments for participants within Questionmark Perception to ensure the delivery environment is suitable for their testing program. Assessments can be scheduled for a specific participant, several specific participants, or groups of participants with conditions specified, such as:

- ► Requiring a participant to log in with a username and password
- ► Taking the assessment once only, a few times, or as many times as the participant likes
- ► Taking the assessment after and before certain dates/times
- ► Taking the assessment only under the watchful eye of a proctor
- ► Accommodating participants with special needs (e.g., visual impairment) by overriding time limits set for closely timed tests
- ► Only at a certified test center
- ► Online or using Questionmark to Go (the disconnected client)

Maintaining up to date information on participants helps ensure both security and accurate reporting of results. The information regarding participants is important for record keeping, results analysis, and security and therefore has strong implications for legal defensibility. Questionmark Perception participant management capabilities enable:

- Manual updating/entering of participant information
- Importing participant information via data files
- Organizing participants properly into groups and hierarchical subgroups, each with designated default test centers, enabling:
    - o simple and secure scheduling of assessments
    - o a streamlined reporting process
    - o reduced possibility of errors in scheduling and reporting

4.3. Delivery

The delivery of an assessment to participants has a great deal to do with legal defensibility. The medium of the delivery method and the security of that method are critical aspects of the assessment process. In addition, the interface through which

participants take the assessment has important implications for fairness and equal access. Perception allows assessment delivery via a number of options:

- ► Standard web browsers
- ► Secure web browsers
- ► CD-ROM and Windows Desktop
- ► Printing and Scanning
- ► PDAs and Smartphones

It is important to choose the delivery options that fit your testing program to ensure legal defensibility.

Questionmark Secure is an essential feature for any organization concerned about legal defensibility in high-stakes assessment environments. Questionmark Secure is a secure player that restricts access to other programs on a computer and focuses the assessment session on the assessment being administered. Providing a secure and "cheat/leak-resistant" environment for all participants is a crucial aspect to the validity of results and fairness to test-takers. Some of the features of Questionmark secure are:

- ► Stops people from printing questions
- ► Stops people from typing in their own URL
- ► Always displays in full screen, making it impossible to maximize or minimize
- ► Stops people starting a new task
- ► Does not provide menu options or icons
- ► Disables control keys and task switching
- ► Disables right-click menu options
- ► Disables the ability to start new applications
- ► Prevents going backwards to a previous page
- ► Stops people exiting in a high-stakes, proctored environment
- ► Hides the HTML source
- ► Provides an API to control certain functions of a browser from the server
- ► Can command Questionmark Secure to display a toolbar

Perception's Save-As-You-Go feature allows assessment answers to be auto-saved as the participant is taking the test, allowing participants to resume an assessment where they left off and reducing test taking frustration in the event of technical failures or other distractions. This assessment delivery capability directly conforms to the ATP Guidelines for Computer-Based Testing (2000).

Questionmark Perception allows two methods of administering assessments to participants: a) question list view and b) question-by-question view.

a) The question list view presents all questions on an assessment on one page in which participants can scroll though all the questions, making changes to questions they have already responded to and submitting their answers. This view is very much like paper-and-pencil-based testing in the sense that participants can go through an assessment booklet answering whichever questions they would like in the order they would like.

b) The question-by-question view allows the assessment administrator to administer one question at a time to the participant, controlling how the participant experiences the assessment and how many question participants see at a time.

Either of these methods of assessment administration can be legally defensible, although each have advantages depending on the purposes of the assessment program and security requirements (e.g., if the assessment administrator would like to limit the number of questions a participant sees at a time to reduce the on-screen exposure of questions).

Questionmark Perception Printing and Scanning allows the printing of assessments from Perception, delivery on paper, scanning of results, and secure upload of results into Perception for use in reporting and electronic record keeping. This capability provides a viable delivery alternative, along with many of the authoring and reporting benefits described in this paper, when computer-based delivery is impractical. If sound administration practices are followed, the process of paper-and-pencil administration—coupled with the electronic storage and reporting/analysis or results—can help ensure legal defensibility.

4.4. Reporting

Reporting is a key area in which organizations can evaluate the performance of participants, and the performance of assessments, to ensure legal defensibility. Reports are where organizations can interrogate the data collected and inform decisions to improve development processes and ensure that the assessments are doing what they were designed to do well.

Questionmark Perception Enterprise Manager offers twelve report types:

- Coaching Report
- Item Analysis
- Question Statistics Report
- Assessment Overview Report
- Survey Report
- Score List Report
- Transcript Report
- Grade Book Report
- Gap Report
- Test Analysis Report (Perception version 4.4+)
- Export to ASCII
- Export to Excel

All of these report types can be used to help evaluate and ensure the legal defensibility of an assessment program and are discussed in this section.

4.4.1. Coaching Report

The Coaching Report provides information on the performance of one participant on one assessment, including the question level details such as the answers that participants have provided. In all stakes environments (although certainly more so in the low and medium stakes) this report can be used to review results with participants to provide one-on-one feedback to guide and facilitate learning. When practice quizzes are administered, the Coaching Report provides participants with information to improve their learning and therefore help ensure that they perform as well as possible on the high-stakes assessment (e.g., exam).

For high-stakes exams, the Coaching Report provides two important benefits with regards to legal defensibility:

1. The report can be used to investigate suspected cheating by comparing answers given by one participant with those of another. This allows an organization to conduct an investigation to determine if answer copying had occurred on multiple question types such as multiple-choice type questions (e.g., if two participants selected a large number of the same *incorrect* answers to questions, this could be a flag that copying had occurred) and essay questions (e.g., if two participants' essay responses are highly similar, this would be a flag that one copied his or her responses from the other participant).
2. The report provides a record of the questions presented and answers given in the event of an appeal or legal challenge. This report is useful in appeal processes in which a participant has appealed his or her score and has requested to see his or her responses to questions on the assessment. Customizing the report makes it possible to hide certain information that the participant should not be seeing (e.g., correct answers for all questions).

4.4.2. Item Analysis Report and Question Statistics Report

Both the Question Statistics Report and the Item Analysis Report provide detailed, classical test theory statistical information on the questions (items) that make up an assessment. This information can be used to determine how well the questions perform psychometrically. The psychometric performance of items is the second necessary piece of evidence (the first being content review validation) that questions are of high quality. High quality items composing a test will help to ensure acceptable test reliability and validity of results because the building blocks of assessments (and therefore assessment results) are items.

The Item Analysis Report and Question Statistics Report provide psychometric information, pertinent to legal defensibility, in two main areas:

a. The difficulty of questions (p-value)

b. How well questions differentiate/discriminate between participants of low/medium/high ability levels (a number of statistics including the Item Total Correlation, Discrimination statistic, and Outcome Information analysis)

The p-value of a question provides information on the proportion of participants who selected a correct answer for a question. It is important to review each question composing an assessment to ensure that the difficulty of the questions are expected (i.e., that the questions are not too easy or too difficult). Questions that are extremely difficult may not be considered appropriate to include on the assessment and may be dropped after a content review. Questions that are extremely easy do not do much in the way of measuring participant ability and therefore are not that useful to include on an assessment.

The Item Discrimination and the Item Total Correlation can help to diagnose problems with questions such as miskeyed questions or ambiguously worded questions. If either of these statistics is close to or below zero, this is a sign that the question has low discrimination and should be examined closely.

Other information that the Question Statistics and Item Analysis Report contain are:

- *The number of participants who missed a question*: If an unexpected number of participants are missing a question (i.e., not responding) this could be a flag that there is a problem with the question. For example, it could be the case that the question is worded ambiguously and so participants are unsure how to answer the question and leave the question blank.
- *The number of responses for each question*: The more participants that answered a question, the more stable the information regarding each question. Therefore, care should be taken when basing decisions on small numbers of responses per item (e.g., statistics based on four people answering a question will not yield very robust results). The more responses to an item, the more certain it is that the statistics generated for that question are accurate and therefore legally defensible.
- *The times that the item is presented*: The more times that a question is presented to participants, the more the question is "exposed." Item exposure is related to question bank security. Item exposure can be tracked with questions retired on rotation to avoid them being seen by too many examinees.

4.4.3. Assessment Overview Report

The Assessment Overview Report provides information on how many participants have taken each assessment, average scores for the assessment, and other information regarding the results. The Assessment Overview Report provides a quick and easy way to interpret overall summary of assessment results. This can help to flag potential problems, which can in turn lead to a more detailed analysis using other reports.

4.4.4. Survey Report

Survey information can be useful in obtaining feedback from participants regarding their experience during or before a medium- or high-stakes assessment. For example, after a medium- or high-stakes examination, consider administering a short survey to obtain feedback from participants. This is an opportunity for participants to raise any issues they may have had regarding the administration environment (flagging potential proctor problems), unfair questions (flagging or preparing for potential appeals of questions), and suspicious behavior (flagging potential cases of cheating). It also provides information on how the participants have been taught in different areas (e.g., regional instructional centers). For example, if several participants at a specific writing center that serviced a specific regional area of participants stated in their survey that they did not cover the material for specific questions from a certain topic, this could be a flag that the instructor was not following the course curriculum or that there were other problems with the teaching process.

4.4.5. Score List Report

The Score List Report provides information on all participants who took one assessment. Like most Perception reports, the Score List Report is customizable to allow for only the fields of interest to be displayed (e.g., question information may not be appropriate to show).

The Score List Report is an excellent resource for instructors because it allows for a snapshot of all the participant results for an assessment in one report. This report enables instructors or administrators to:

- Determine whether the test performed as expected for participants (e.g., did the participants who were at the top of the class throughout the course do well, as expected?), which provides valuable information on the validity of the assessment.
- Review the IP addresses of the participants who took the assessment to ensure that all participants took the assessment from an expected IP address. This is important for security and therefore legal defensibility.
- Track the progress of participants as they are taking the assessment, showing which participants are in progress and which of them are finished.
- Ensure that all participants completed the assessment in an expected timeframe. If a few participants did very well on the assessment but only took a fraction of the time it took other participants, this may be a flag that something suspicious was going on. If many participants were not able to finish the assessment, this may flag an issue of fairness (if the assessment was designed so that most participants should have time to finish answering all questions, then in the report one should see evidence that most participants were able to finish).

4.4.6. Transcript Report

The Transcript Report shows the results for one participant on multiple assessments. This provides an excellent mechanism to:

- Archive participant results, necessary to ensure legal defensibility of the assessment program.
- Review historical performance if a participant is suspected of engaging in answer copying or other forms of cheating.

4.4.7. Grade Book Report

The Grade Book Report allows users to produce a grade book composed of multiple assessments for groups of participants. The report allows users to weigh assessment results to calculate a weighted final score for participants.

Many organizations administer multiple assessments to participants and often these assessments are taken together to determine whether a participant has met certain knowledge/skill criteria. The Grade Book Report provides the venue to produce appropriate and fair weightings of assessments and keep a permanent record of these decisions.

4.4.8. Gap Report

The Gap Report allows the comparison of different groups of participants, or the same group of participants but at different times. The report allows the investigation of differences achieved at a question, topic, or assessment level.

The Gap Report can be used to determine if an assessment performs equitably for different demographic groups, such as males and females. Investigating differences on assessment performance between groups of participants can determine whether bias is present and can have an impact on the use, and therefore fairness, of assessment results. Gender differences are an area where legal issues can arise, so ensuring that an assessment performs similarly for both genders is an important piece of evidence in the overall suite of evidence that an assessment, and broader assessment program, is legally defensible.

4.4.9. Test Analysis Report

The Test Analysis Report is designed to give a comprehensive summary of the performance of an assessment to evaluate and ensure legal defensibility. The Test Analysis Report provides detailed statistical information regarding the performance of

an assessment and can be used for exams, tests, and quizzes. The Test Analysis Report provides a detailed table of statistics, including a very important Cronbach's Alpha reliability coefficient, a frequency distribution of total test scores, and a histogram of the total test scores. Starting with the table of test statistics, it is possible to determine whether the assessment performed the way that it was expected to perform, in terms of the central tendency (e.g., mean), variation (standard deviation), and reliability measure, which will be discussed in more detail below. The frequency distribution and histogram allow graphical diagnosis of potential problems, such as whether groups of participants clustered together and obtained similar scores, which could be a flag for answer copying (e.g., if a large number of participants obtained exactly the same score of 100%, this could be a sign that the assessment questions had been leaked prior to the administration date).

A key piece of information provided in the Test Analysis Report is the Cronbach's Alpha reliability coefficient. This provides an internal consistency reliability value for the assessment and is a measure of the statistical reliability of the assessment score—the degree to which questions composing the assessment "hang together" to measure the same construct. Chapter 3 discussed acceptable ranges of Cronbach's Alpha for high-stakes testing scenarios as well as tips on what you can do to improve the reliability of an assessment. In general, for a test that measures a single construct, Cronbach's Alpha values above 0.90 are considered acceptable for high stakes assessments, and values between approximately 0.70 and 0.89 are considered acceptable for medium-stakes assessments. The reliability coefficient of an assessment is one very important piece of information that will be required in order to demonstrate that an assessment is legally defensible (a test must demonstrate acceptable reliability in order to be considered valid).

Reliability should be as close to +1 as possible not only on the overall assessment but also on each topic (depending on how the test is designed). This is why summary information, including Cronbach's Alpha, is displayed at both the overall assessment and topic levels. If the test is designed to measure dimensions (topics) within an overall construct, the Cronbach's Alpha values for those topics are important. The overall reliability will be reduced if an assessment is composed of multiple topics that measure very different constructs.

4.4.10. Export to ASCII and Export to Excel

The Export to ASCII and Export to Excel features allow users to format the data that they need to export the result from Perception to other systems. The exported information can be stored in an organization's central records system for archiving purposes. Organizations may wish to conduct further advanced psychometric analyses on results in order to investigate specific aspects of legal defensibility issues.

4.4.11. Using Filtering in Questionmark Reports

An important feature in Questionmark Perception reporting is the ability to use filters in the report generating process. This allows organizations to produce reports that are broken down by date, use special (demographic fields) to investigate demographic differences, or filter on groups and examine participant group differences. For example, if one wanted to determine how an aptitude test performed for different ranks of military officers, one of the special fields could be used to hold the military ranks of participants who took the assessment. This special field could then be used as a filter to produce separate Test Analysis Reports, one for each rank.

Determining how assessments perform over time and across different groups or demographics is highly valuable information in the quest to ensure the legal defensibility (validity investigation) of an assessment program and Questionmark Perception report filters can help toward this end.

About Questionmark:

Questionmark has been producing testing and assessment software since 1988. Businesses, governments, schools, colleges, and universities in more than 50 countries use Questionmark software. Questionmark has more than 2,500 customers using Perception, with approximately 14,000 authoring systems installed, and systems shipped have been configured to serve millions of participants. Typical applications include exams, quizzes, study aids, course evaluations, surveys, diagnostic tests, pre-course skills assessments, and course evaluations.

Questionmark Offices

Question*mark*
535 Connecticut Avenue
Suite 100
Norwallk, CT 06854
Tel: (800) 863-3950
(203) 425-2400
Fax: (800) 339-3944
info@questionmark.com

Question*mark*
4th Floor, Hill House
Highgate Hill
London N19 5NA
United Kingdom
Tel: +44 (0)20 7263 7575
Fax: +44 (0)20 7263 7555
info@questionmark.co.uk

Question*mark*
Parc d'Activités Economiques
Avenue Léon Champagne, 2
B – 1480 Tubize (Saintes)
Belgium
Tel: +32 (0)2 398 02 01
Fax: +32 (0)2 398 02 00
info@questionmark.be