Pearson VUE

Developing ***valid, fair*** *and* ***reliable*** *exams*

Some high-stakes examinations are created to distinguish test takers who demonstrate required professional competencies from others who do not. Other examinations place test takers along a continuum so that valid comparisons can be made. Whatever an examination's purpose, to achieve it, the exam must be valid, fair, and reliable.

Validity, simply put, is concerned with answering two questions:

1. Does the test measure what it is intended to measure?

2. Are the interpretations drawn from the test scores appropriate and justifiable?

Fairness is a specific validity issue. The examination should assess what it is designed to measure without the influence of extraneous factors. Results should not be affected by the specific version of the exam a test taker happened to receive. Nor should test taker demographics, such as gender, ethnicity or disability status, interfere with an assessment of the ability the exam is intended to assess.

Reliability concerns the consistency and stability of the measurement. Are the test results reproducible? In other words, if the test taker were to take the examination again without any changes to the test taker's circumstances (such as further study or revision), would the test results be the same? Exam reliability may also be influenced by the conditions in which the exam is administered. For instance, a noisy environment may cause test takers to make errors in responding to test questions.

The quality of the test items (questions) and how they are worded can affect the reliability of the results as well as exam fairness. Items which are ambiguously worded may mean that test takers do not respond the way they would if the intent of the item was clearer. Test items containing language that is unfamiliar to certain groups of test takers is also a fairness concern.

These three concepts underlie all aspects of developing, constructing, and administering high-stakes exams. [1]The process for defining what the exam will assess and [2]item writing and review procedures contribute to content - and construct-oriented validity.

[1]Test Development Guide 1: Creating test blueprints
[2]Test Development Guide 2: Developing high-quality test items

Developing valid, fair and reliable exams

Test administration models

The process for constructing exams – determining which items from the item bank will be administered to test takers – is instrumental to providing valid, fair and reliable examinations. Each version of the examination should be constructed according to the test blueprint with both content and statistical specifications addressed by the test administration model. Content specifications concern the distribution of questions across content or performance areas. Statistical specifications deal with item difficulty and statistical equivalence. All versions should be constructed to the same test specifications so that it does not matter which version of the exam each test taker receives – it will be comparable to the version others sit.

If test takers are to receive their results immediately (one of the advantages of computer-based testing), then it is strongly recommended that only items whose statistical properties are known are used in determining test taker performance.
Pre-testing of items can occur in a number of ways, including seeding unscored items amongst the scored items in an exam.

Test administration models supported by computer-based testing, as outlined in Becker & Bergstrom (2013), include:

Fixed linear forms:

This test administration model most closely resembles pencil-and-paper testing. A set number of alternative versions (or forms) of the test are developed prior to administration of the test.
All test takers who receive a given test form are administered the same set of items; in computer-based testing, the order in which the items are administered is frequently randomised. The alternate forms should be built to the same content specifications and they should be statistically equivalent so that no one form is harder or easier than the others (in other words, the forms are "parallel" to each other).

Glossary of terms

Equating:
The process of statistically adjusting the scoring of alternate forms of a test so that the scores on different forms are expressed on the same scale. Equating is performed to address minor differences in difficulty across the alternate forms.

Item Response Theory (IRT):
A statistical model for analysing and scoring tests that is based upon the concept that the probability of a correct response on any test item is a function of person and item characteristics; the relationship between these characteristics and the probability of a correct response is modelled by an item characteristic curve (ICC).

Pre-test items:
Newly written items that have not yet been made operational. They are administered to test takers for the purpose of collecting data about the items (i.e., for computing item statistics). Pre-test items may be presented as unscored items amidst scored items in a test or in a separate test referred to as a beta test.

Psychometrician:
An expert in the theory and practice of measurement who typically has an advanced graduate degree from a university, usually from an educational measurement programme or quantitative psychology programme.

Linear-on-the-fly testing (LOFT):

In a LOFT exam, test items are selected for administration to individual test takers based upon pre-determined content and statistical constraints so that test takers receive comparable parallel test forms. It is called "on-the-fly" testing because intact test forms are not developed prior to testing; rather, the items for an individual test taker are selected when he or she sits the exam. It is considered a form of "linear" testing because the selection of items does not depend on the test taker's performance on previous test questions. LOFT increases test security by limiting the exposure of all test questions since each test taker receives one of a large number of possible parallel forms.

Computer-adaptive testing (CAT):

Successive test items are selected to be administered to test takers from a pool of questions based upon the test taker's performance on previous questions – with a more difficult question being selected after a correct response, and an easier question being selected after an incorrect response. A computer-adaptive test provides high-quality measurement and is more efficient than traditional linear testing models because test takers are not administered items that are too easy or difficult for them. Adaptive tests rely on statistics based in Item Response Theory (IRT) for scoring and question selection.

Computer-adaptive multi-stage testing (MST):

Multi-stage testing is similar to CAT in that test taker performance on previous questions determines which questions are seen next by the test taker. Unlike CAT, MST administers sets of questions in modules. Therefore, sets of questions (rather than individual questions) are selected for administration based upon the performance of the test taker on previous questions.

The appropriate test administration model for your examination is dependent upon a number of factors, including test taker volumes, the size of the item bank needed to support the model, reporting requirements, and how important it is for you to review intact test forms prior to administration.

How Pearson VUE can help

Pearson VUE's Measurement Services team can assist you with the entire psychometric scope of work needed to support a high-stakes examination programme. Our psychometricians work with clients on test design, exam and item analyses, test construction, equating and scaling, and setting the standard (pass point). We can help you determine the test administration model most appropriate for delivering your examinations.

Our Measurement Services team includes approximately 25 PhD-level psychometricians with over 200 years combined experience. As psychometric resources are limited and highly valued within the testing industry, Pearson VUE's assets in this area are significant.

Reference

Becker, Kirk A. & Bergstrom, Betty A. (2013). Test administration models. Practical Assessment, Research & Evaluation, 18(14). Available online: http://pareonline.net/getvn.asp?v=18&n=14.

To ***learn more*** or ***talk to us***
visit pearsonvue.com