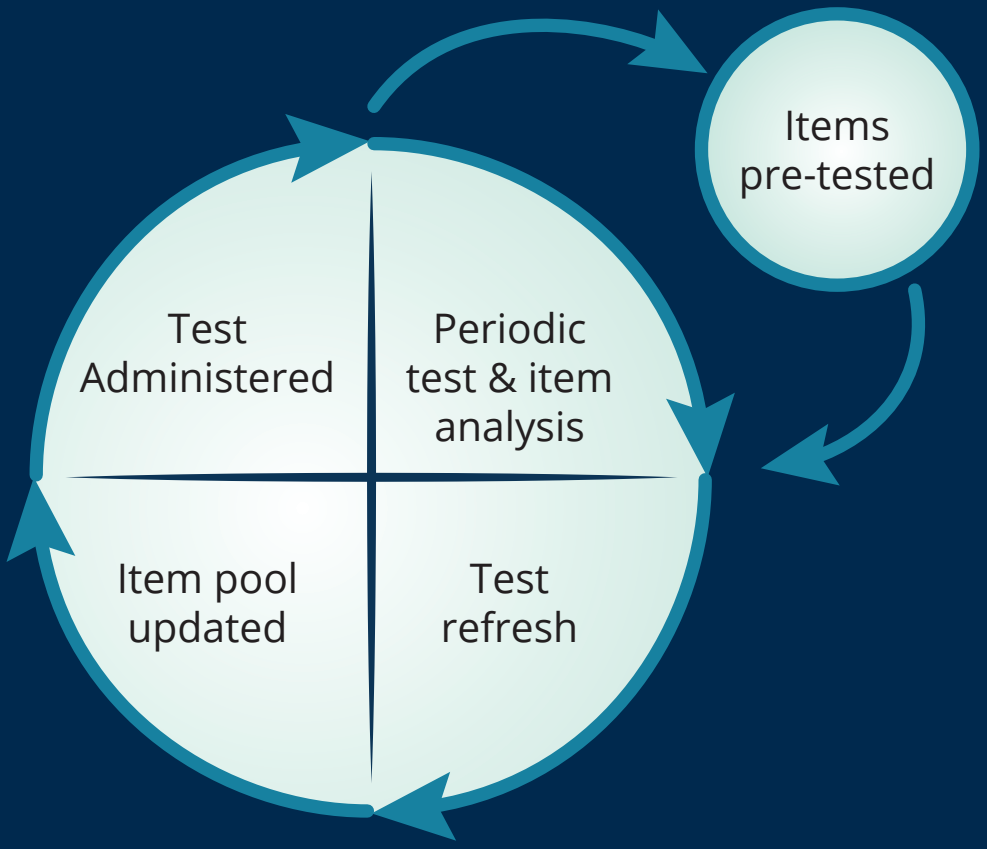
Pearson VUE

Evaluating performance



Assessment is essentially a data collection exercise. Examinations are instruments for collecting information about particular test taker traits and abilities in order to make decisions. Does the test taker meet competency standards to practice a profession? How does the test taker compare to others who are applying for admission to an academic study programme? The data collected through examinations are not only used to evaluate test taker performance but also how well the examination itself, and the items within it, are performing.

Psychometric analysis

Pre-testing

Pre-testing (or trialing) of test items refers to the administration of the items solely to gather performance statistics to determine if the items should be included in the operational item pool from which items are selected for future administration and scoring. Pre-test items do not count toward the test takers' scores. Pre-testing of test items is especially recommended if you wish to provide immediate score reporting after exam administration.

For new examinations, pre-testing may be conducted by administering a beta test. Beta tests aren't scored; they are conducted solely for the purpose of collecting item performance statistics. Moving forward, items may be trialled by embedding or seeding a small number of unscored, pre-test items within the operational, scored items in a live exam administration.

Item analysis of operational items

Item performance may change over time.

Therefore, psychometric analysis of operational scored items should be conducted to evaluate which items should remain in the operational pool and which items should be retired from use. Psychometric analysis of the operational items should be conducted after each exam administration or periodically in the event of continuous on-demand testing.

Exam analysis

In addition to an item-level analysis, psychometric analysis should also be performed at the exam level, and reliability a primary focus. Other performance data typically monitored at exam-level includes summary statistics on test taker performance and exam timing information and with reliability the primary focus.

One of the advantages of computer-based testing is that timing data is automatically captured. Basic analysis of exams, such as the mean time taken to complete the exam is routinely calculated.

More detailed analyses may be routinely performed depending upon the type of examination, the scoring methodology and other circumstances of the examination program. More analysis may also be performed as part of a special study to address a particular issue, for instance, if there are concerns about the number of test takers who are unable to complete the exam, a more detailed timing analysis may be performed.

Exam-level performance data:

- Reliability
- Descriptive statistics on test scores
- Number of test takers
- Pass rate
- Mean score
- Standard deviation
- Descriptive statistics on exam time
- Mean time taken on exam
- Percentage of test takers using the entire testing time

Standard setting

Standard setting is the process of determining the performance standards that must be achieved to receive a particular designation (for instance, the pass mark for a pass/fail examination). Unless there is a rationale for using a norm-referenced standard, criterion-referenced standards are the most defensible and is widely recognised as the method of choice for high-stakes examinations, such as licensure or certification exams.

There are a number of well-recognised standard setting methods. Most of these methods involve the use of subject matter experts to review individual test items and candidate performance data on these items, if available. Information gained during this review is used to set the performance standards or pass mark.

Glossary of terms

Beta test:
A test administered for the purpose of collecting information about individual test items, to determine their effectiveness and provide data for determining whether they should be included in future tests. Beta tests are typically not scored at the time of administration; they may be scored after the item statistics have been computed, or they may never be scored.

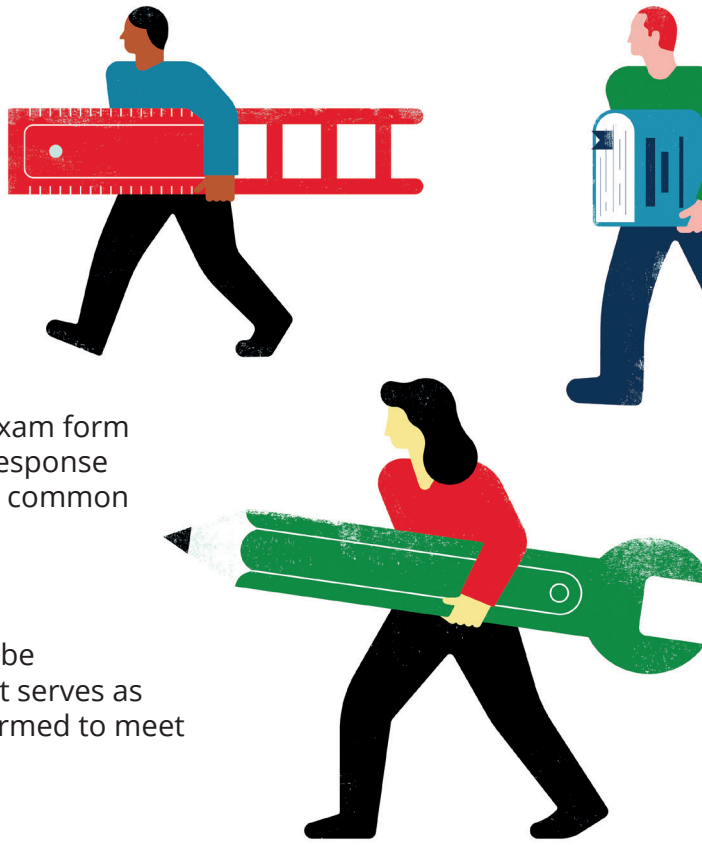
Classical test theory (CTT):
An exam development and evaluation framework derived from the premise that any test score can be expressed as a combination of two components: (1) a 'true' score, or a test taker's 'true' status on the construct measured by the test, and (2) random error.

Criterion-referenced test:
On a criterion-referenced test, each test taker's performance is compared to a pre-determined criterion or standard. The performance of other test takers does not influence the score of individual test takers.

Item analysis:
Statistical investigation of the performance of test items to obtain information about the quality of the items.

Equating and scaling

The goal of the equating process is to make certain that scores are comparable across different test forms of the examination so that it doesn't matter to test takers which form of the examination they are given.



Because the raw scores on different versions of the same examination do not reflect an equivalent level of underlying knowledge or ability, different forms of an examination are scaled to have the same passing scaled score following the equating exercise. Through the exercise, a given scaled score always reflects the same performance level or underlying ability, regardless of the exam form that was taken. Either Classical Test Theory (CTT) or Item Response Theory (IRT) can be used to quantitatively link all items to a common benchmark scale.

Technical reporting

All relevant psychometric data for a given testing cycle can be captured in a detailed technical report. The technical report serves as documentary evidence of the psychometric activities performed to meet organisational accreditation and other requirements.

How Pearson VUE can help

The Pearson VUE Measurement Team offers a full suite of psychometric services to assist you with the design, development and analysis of your examinations. We have over 25 psychometricians with advanced degrees in measurement who are able to advise you on the most appropriate standard setting methodology and conduct standard setting exercises. We can also advise on the most appropriate pre-test strategy.

Pearson VUE psychometricians routinely prepare detailed technical reports summarising candidate, exam and item performance and can make recommendations for examination program maintenance and improvement.

Pearson VUE psychometricians are adept at both Classical Test Theory (CTT) and Item Response Theory (IRT) analyses.

Item discrimination:

A measure of how well an item differentiates among test takers according to their skill or proficiency in the content being measured.

Item Response Theory:

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.

Norm-referenced test:

A method of scoring in which the test taker's performance is compared to other test takers (called the "norm group") and the score reflects the test taker's relative position in that group. Therefore, the score is not based upon a pre-defined criterion for mastery of the subject area in question but instead is an indication of how well the test taker did in comparison with others.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.(1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
International Test Commission (2001). International Guidelines for Test Use, http://www.intest.com/test_use.htm (Retrieved 7 January 2012).

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

PV/4 Eval/UK&APAC/8-17