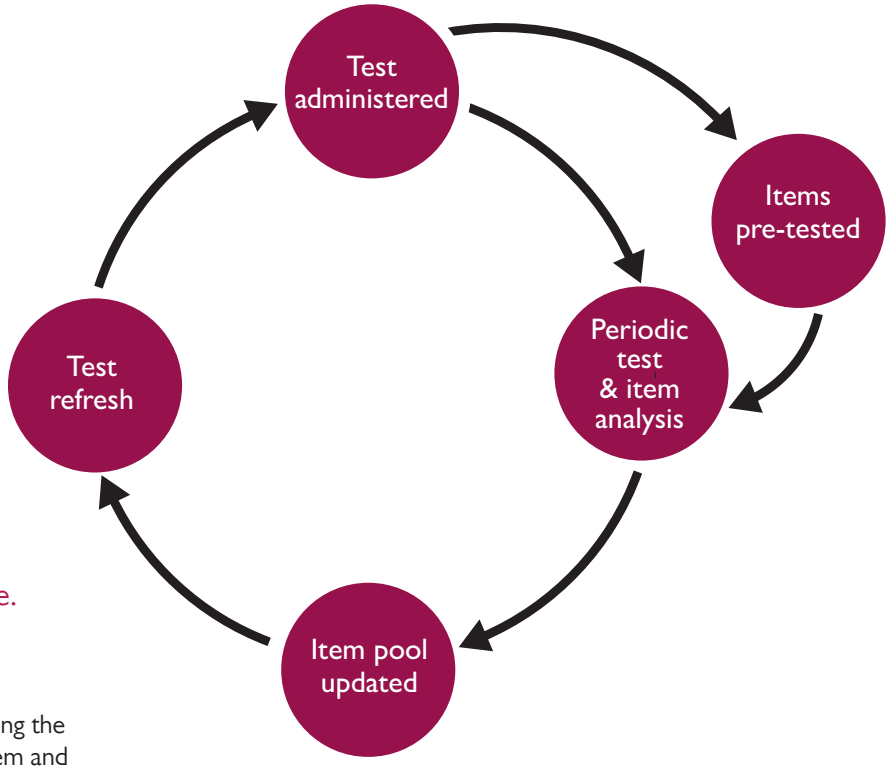
# Assessing and maintaining test quality




In an on-demand computer-based testing (CBT) environment, the process of item and exam development is a continuous cycle.

Items are pre-tested by seeding un-scored items among the scored items in the tests administered. Periodically item and test data are comprehensively analysed. These analyses serve to ensure that acceptable standards of quality are maintained and to inform those responsible for content development about item performance and assist them in making decisions on the disposition of items. It is in this manner that the item bank is continually replenished so that the test can also be periodically updated.

## Psychometric analysis

To make sure that test items meet acceptable standards of quality, Pearson VUE psychometricians continually monitor performance of examinations and items. This includes a regular review of pass/fail rates and review of item statistics for parameter drift, an indication that the item has gotten easier or harder over time. Among the types of test and item data monitored are
the following:

Standard error of measurement
- Number of respondents
- Number of respondents answering correctly
- Item difficulty
- Standard Error of Measurement
- Item discrimination
- Response time on item

Response option (distractor analysis)
- Number of respondents
- Number of respondents answering correctly
- Response difficulty
- Response discrimination

Test/item pool
- Distribution of the items by content
- Distribution of the items by difficulty
- Average and range of difficulty by content area

Pearson VUE psychometricians prepare a regular technical report to inform the test sponsor and appropriate stakeholders on test and item performance. Typically, the technical report contains item level information as well as information about the cut score, reliability, raw and scaled scores, and the procedures used for equating. The following sections discuss some of the information contained in technical reports.

# Assessing and maintaining test quality

## Reliability and standard error of measurement

Measurement error is always present in test scores. This does not mean an error has been made in preparing or scoring the test, but rather that the test-takers' scores are not perfectly consistent either on repeat administrations of the same form, from one test form to another, or from one administration to another. An estimate of the extent to which scores on a test are free from measurement error is captured in a coefficient called reliability. The reliability coefficient and the standard error of measurement are the two most commonly used indicators for conveying the accuracy of test scores.

Various types of reliability coefficients exist, each designed to estimate different types of measurement error. One of the most commonly used coefficients is referred to as internal consistency reliability. Internal consistency reliability coefficients estimate error stemming from the collection of items in a specific form and the way test-takers respond in a single administration. Other, less commonly used types may assess changes in test-taker's performance over time or variation across test forms.

Reliability coefficients can range from .00 to 1.0 and generally fall between .70 and .95. The closer the reliability coefficient is to the maximum, the less the measurement error in test scores. Test characteristics that can influence internal consistency reliability measures are test length (number of questions), time allowed and relationships among the items. Generally, longer tests, those that allow sufficient time to respond to all questions, and tests composed of items that all measure a single, unitary construct or trait will have higher levels of internal consistency reliability. These factors should be taken into account in interpreting reliability or comparing values across different tests.

The standard error of measurement (SEM) is an estimate of how far off a score is from a test-taker's actual or "true" proficiency as a result of the measurement error discussed earlier. To understand the meaning of the SEM, consider a hypothetical situation in which a test-taker with a given level of true proficiency in mathematics takes a form of a maths test many times, without memory of the questions from time to time and without a change in his or her true proficiency. It is very unlikely that the test-taker will obtain exactly the same scores. Instead, the scores will probably be close to each other and close to his or her true proficiency. This variation in scores – the differences between observed and actual proficiency – is what is captured by the SEM. Specifically, the SEM is the standard deviation of the distribution of these differences. SEM and reliability are inversely related to each other; the higher a test's reliability coefficient, the smaller the SEM, and vice versa.

For any given test-taker, it is not possible to determine how much his or her observed score differs from their true proficiency. However, measurement theory provides a basis for using SEMs to establish score bands or confidence bands around a candidate's score. Under certain, usually reasonable, assumptions about measurement error, a test-taker's observed score will be within ±1 SEM of his or her true proficiency about 68% of the time and within two SEMs about 95% of the time.

## Item analysis

At a minimum, item analysis using Classical Test Theory (CTT) should be performed and reported. The two main item indices of CTT are difficulty and discrimination. Items which should be reviewed by content experts are flagged using the item analysis results.

## Glossary of terms

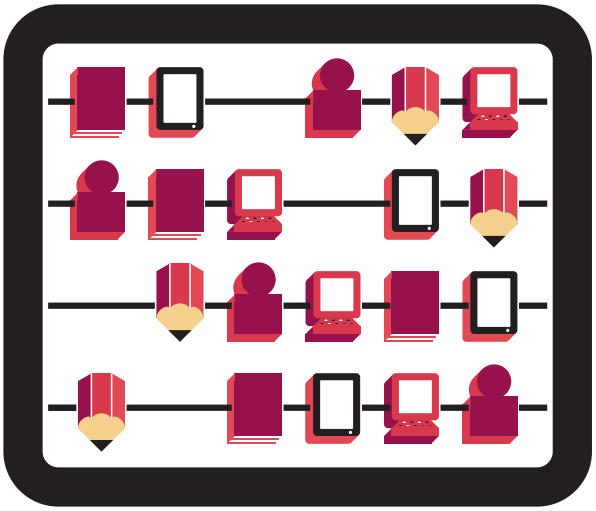
Cronbach's alpha
A statistic that is used to estimate the reliability of scores on a test. What alpha measures is internal consistency – the extent to which the items measure the same knowledge or skill. Under some assumptions that are usually reasonable, alpha also indicates the extent to which test-takers would perform similarly on two different forms of the same test.

Distractor analysis
A statistical evaluation of how well the distractors on a selected-response type of item perform.

Difficulty
A measure of how hard an item is. Also called the facility or p-value in Classical Test Theory.

Discrimination
The relationship between success (as defined by answering correctly) on a particular item and success on the test as a whole.

Item analysis
Statistical analysis of test-takers' responses to test questions, done for the purpose of gaining information about the quality of the test questions.



Item difficulty in CTT is simply the proportion of test-takers who answer the item correctly. The resulting statistic, referred to as the p-value or the facility value, has values which range from 0 to 1.0; the higher the value, the easier the item. While there may be a legitimate reason for including a very easy or very hard item on a test, four-option multiple-choice items with p-values below 0.20 or greater than 0.95 may be problematic. Generally, a test should include items with a range of p-values; too many easy or too many hard items on the test are not desirable.

Item discrimination is usually computed as a point-biserial correlation, the correlation coefficient between the item score and the total test score. The point-biserial correlation indicates the degree of relationship between item performance and test performance. The resulting statistic ranges from -1.0 to 1.0. A high positive correlation (for example, greater than .20) is desirable. A negative correlation means that test-takers who perform less well on the test overall perform better on the item.

Depending on the test purpose and the testing programme, other item analyses, such as Item Response Theory (IRT) and differential item functioning (DIF), may be performed and reported. IRT is a mathematical model for describing the relationship between test-takers' performance on an item and their ability level. DIF is "a statistical property of a test item in which different groups of test-takers who have the same total test score have different average item scores or, in some cases, different rates of choosing various answer options" (AERA, APA, NCME, 1999, p. 175).

## Content review

Items flagged for their statistical properties are reviewed by content experts. Content experts make diagnostic assessments on the items and determine their disposition. The following table outlines potential causes of poorly performing items.

| Item Statistical Property | |
|---|---|
| Low p-value | Negative or low point-biserial correlation |
| Potential cause: | |
| Key is incorrect<br>There is more than one correct answer<br>Item contains content that is rare or trivial<br>Item is ambiguous | Key is incorrect<br>There is more than one correct answer<br>Item is too difficult and test-takers are guessing<br>Item is ambiguous<br>Item is testing something different from the other items |

Following the content review, each item is either: 1) approved for continued use, 2) sent for re-write and then pre-test, 3) retired and archived in the item bank.

## Item banking

Use of an item bank enables the tracking of items, provides an historical record of item performance and allows querying by item characteristics including statistical properties. Pearson VUE Content Development staff monitor the item inventory in the item bank including the number of items by test blueprint category and difficulty level. A gap analysis of the item bank is used to target item writing activities.

A complete maintenance programme of statistical analysis, content review of item performance and targeted item writing activities ensures that test quality is sustained.

Kuder-Richardson Formula 20 (KR20)
Like Cronbach's alpha, a measure of internal consistency reliability. Unlike Cronbach's alpha, the KR20 can only be used with dichotomous measures.

Reliability
A measure of the consistency or stability of test scores for a group of test-takers over time, administration conditions, test forms, or samples of items.

Standard error of measurement (SEM)
The standard error of measurement reflects the amount of random error in a set of test scores – i.e., it is a measure of the tendency of test-takers' scores to vary because of random factors, such as the particular selection of items on the form the test-taker happened to take. The smaller the SEM, the smaller the influence of these factors.

### 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).

## Pearson VUE Sales Offices

### Americas


Chicago, IL
+01 888 627 7357
pvamericassales@pearson.com

*Global Headquarters*
Minneapolis, MN
+01 888 627 7357
pvamericassales@pearson.com

Philadelphia, PA
+01 610 617 9300
pvamericassales@pearson.com


### Europe, Middle East & Africa


Dubai, United Arab Emirates
+971 44 535300
vuemarketing@pearson.com

London, United Kingdom
+44 0 207 775 6737
vuemarketing@pearson.com

Manchester, United Kingdom
+44 0 161 855 7000
vuemarketing@pearson.com


### Asia Pacific


Beijing, China
+86 10 5989 2600
pvchinasales@pearson.com

Delhi, India
+91 120 4001600
pvindiabusiness@pearson.com

Melbourne, Australia
+61 3 9811 2400
pvseasiasales@pearson.com

Tokyo, Japan
+81 3 6891 0500
pvjpsales@pearson.com


## Continually assessing and maintaining test quality

To learn more, visit www.pearsonvue.com

PV/4 On Maint/UK&APAC/1-15