Validity

Written by Barry O'Sullivan

People often talk about a test being "good" or "bad" or whether it is "fit for purpose". In technical language, they are actually talking about validity. Test developers often claim that their test is valid or that it's been validated. But what is actually meant by the term valid? In fact, there are two camps here. One argues that validity is a feature of a test – the test does what it claims to do. The other claims that it is a feature of the decisions that are made based on test performance.

Let's think about those two positions for a moment. If validity is a feature of a test, then we only need to gather evidence that the test is measuring a particular trait or ability. If validity is a feature of the decisions, then the scope of the information required to demonstrate validity is far broader. It should include the same evidence as is expected already, plus evidence that the test is working appropriately within specific educational and social contexts and with specific test takers. Since the latter understanding is the most commonly accepted, let's stick with it.

Sticking with it brings up two very important points: The first is that a test itself is not valid, as validity is related to test use. The next is that validity cannot be just a number. Actually it's an argument build around a set of evidence, which is expected to support any decisions that are made based on test performance.

Even before the birth of the testing industry early in the 20th century, people were beginning to think about the quality of tests and by the 1950s the modern concept of validity was emerging. At that time, the feeling was that any one of three types of evidence was required. These were **construct** (the ability being tested), **content** (what the test contained) and **criterion** (how the test outcome compared to other measures of the same skill, for example from another test or from a teacher).

By the 1990s it was widely agreed that a unitary approach was needed, with evidence from a variety of sources expected to contribute to a single validation argument. By then the types of evidence expected had grown so complex that while the approach became the theoretical norm in the 21st Century, it has never been possible to fully apply it.

Around the turn of the century, the socio-cognitive approach emerged. This approach built on the unitary approach, but attempted to balance the social and cognitive aspects of language ability while making clear how the different types of evidence fitted together to form a fuller picture of validity.

The approach asks that evidence be gathered from the beginning of the development process, focusing on the test taker, the test system (the questions and activities included in the test as well as things such as timing) and the scoring system (including all aspects of scoring and awarding a grade or mark). Evidence should also come from test stakeholders. Stakeholders are those people affected by a test and can include test-takers, parents, teachers, local education officials, policy makers and others.

Once all the evidence has been gathered it must be put together to form a convincing argument to support test related decisions. The most important thing is that a logical and comprehensive set of evidence is presented in an appropriate ways for all stakeholders to understand.

Before we finish, I should mention reliability. You may have heard the term and wonder what it means. Well here's a brief explanation. There are basically two different ways to think about reliability. The first relates to what we call the internal consistency of a test. The second relates to the accuracy and consistency of marking in tests of writing or speaking.

Reliability is usually reported on a scale of 0 to 1 and the higher the score the more consistent the test. One important thing to remember is that the reliability estimate does not tell us that the test is good. It just tells us the questions in the test are consistently measuring the same thing – it's possible that it's consistently wrong!

It was once thought that test developers had to achieve a balance between reliability and validity. These days, reliability is considered to be an aspect of validity and forms part of the argument that the scoring system is working well. It's important to remember that a test itself cannot be valid. Instead, we build an argument based on evidence gathered to support the use of the test for making specific decisions about specific test takers in specific situations. So, while a test may be validated for use in one situation, this does not necessarily mean that it can be used in other situations without additional evidence.

Unfortunately, the cost of gathering and presenting the validity evidence means that many test developers never get round to doing it – so, in some ways, the concept of validity remains more of an aspiration than a reality.