A hitchhiker's guide to validity

Charles Darr

Validity and reliability are two words often associated with assessment, particularly with testing. But what do they mean and how can they inform the decisions we make about assessment as classroom practitioners and school leaders? In the next two editions of Assessment News we take a look at these two terms, starting this time with a discussion of validity. What is discussed below should be considered as a kind of "hitchhikers' guide", as of course there is a lot more than can be said about this very important topic.

What is validity and is it important?

In the past, validity has often been treated as the degree to which a test or assessment tool measures what it claims to measure, as if this were a something inherent in the assessment instrument itself. More recently, however, assessment specialists have argued that validity should not be considered as a fixed property of an assessment instrument. Instead, they propose that validity is better understood as an evaluation of the quality of the interpretations and decisions that are made on the basis of an assessment result—that is, how well the inferences we make or actions we take on the basis of an assessment result can be justified.

Validity can be considered as the key issue in assessment. If an assessment is to going to have any use at all, it is crucial that the inferences and decisions we make on the basis of assessment results are well founded. So, how do we go about judging validity?

Judging validity

Judging validity cannot be reduced to a simple technical procedure. Nor is validity something that can be measured on an absolute scale. The validity of an assessment pertains to particular inferences and decisions made for a specific group of students.

Determining validity, then, involves amassing evidence that supports these interpretations and decisions. The strength of that evidence will lead us to a strong, moderate, or weak case for validity. What evidence we collect will depend on the kind of interpretations and decisions we want to make. The checklist in Figure 1 provides some places we can begin to look for this kind of evidence.

Investigating validity like this has sometimes been referred to as developing a validity argument. Two assessment experts, Robert Linn and David Millar (2005), propose four major considerations that arguments concerning validity should take into account. Three of them—content considerations, construct considerations, and criterion relationships—have traditionally been part of the validity landscape. The fourth one, which has been added more recently but is just as important, is consequential considerations. Below, I have outlined what these considerations entail and how they can help us evaluate validity.

- Do the tasks match the learning intentions we are interested in?
- Does the test cover a wide enough range of content?
- Are there enough items or tasks to cover the scope of what is being assessed?
- Do the tasks require use of the desired skills and reasoning processes?
- Is there an emphasis on deep, rather than surface knowledge?
- Are the directions for the assessment task clear?
- Are the questions unambiguous?
- Are the time limits sufficient?
- Do the tasks avoid favouring groups of students more likely to have useful background knowledge—for instance, boys or girls?
- Is the language used suitable?
- Are the reading demands fair?

Content considerations

When we assess students we can't test everything. It is important therefore, that what is tested is a fair sample of the area of learning we are interested in. Considering the content of our assessments as part of a validity argument involves evaluating how well our assessment tasks represent or sample the learning domain in question. This means that we have to be very aware of our initial learning intentions and able to demonstrate the links between them and the tasks or assessment items we are using.

Content issues should be carefully considered when developing assessment tasks. Sometimes this might involve spending time writing content specifications before designing or choosing the items and tasks to match these. When standardised tests and high-stakes assessments are being written, the developers will often bring in subject specialists, including panels of teachers, to check that both the content specifications and the matching assessment items reflect what is commonly being taught.

As teachers, we also need to examine to what degree the assessment tools we develop (as well as the ones that come pre-packaged) represent the emphasis and scope of the learning domain. Part of this could involve conferring with colleagues and students to evaluate our choices. When an assessment tool provides a fair representation of the learning we are interested in, we increase our ability to make

FIGURE 1. VALIDITY CHECKLIST

- ☐ Do the tasks match the learning intentions we are interested in?
- ☐ Does the test cover a wide enough range of content?
- ☐ Are there enough items or tasks to cover the scope of what is being assessed?
- ☐ Do the tasks require use of the desired skills and reasoning processes?
- ☐ Is there an emphasis on deep, rather than surface knowledge?
- ☐ Are the directions for the assessment task clear?
- ☐ Are the questions unambiguous?
- ☐ Are the time limits sufficient?
- ☐ Do the tasks avoid favouring groups of students more likely to have useful background knowledge—for instance, boys or girls?
- ☐ Is the language used suitable?
- ☐ Are the reading demands fair?

valid inferences about achievement in that learning domain.

Construct considerations

One of the goals of much assessment is to establish whether certain characteristics or traits have been developed. Construct considerations are a way of looking at this. Constructs are specific psychological characteristics or traits, such as a type of reasoning or thinking, that we are interested in assessing. In a mathematics assessment a trait could be "problem-solving skill" or "mathematical reasoning". In a reading assessment it is often "comprehension". When we look at construct considerations as part of a validity argument, we are examining the extent to which the assessment result can be used to make inferences about the existence of a certain construct or constructs. Since assessments are usually focused on particular traits or characteristics, construct considerations are often given the highest priority when evaluating validity.

To make *valid* comments about a construct based on an assessment result, we need to be able to show that the construct is essential for success in the assessment tasks. When important aspects of the construct are under-represented in the assessment, or other factors not related to the main construct (ancillary factors) are required, then the inferences we make regarding the construct will have low validity.

An example of an ancillary factor could be the reading demands of a mathematics test. Unfair interpretations of mathematical ability could be made if the reading demands of a question present extra obstacles, especially for slow or less able readers. Here we have to ask what is really being assessed: mathematics ability or reading comprehension?

In a similar way, some assessments might allow us to make valid inferences about the levels of achievement on a construct for some students, but not others. For example, an assessment task that involves a series of computational questions might allow valid inferences to be made about the mathematical problem-solving ability of younger students, who will often have to apply problem-solving strategies to find an answer. However, inferences about problem solving might not be so valid for older students, who often apply methods they have memorised and no longer reason their way to a solution.

Considering the treatment of the construct in an assessment task means that we have to have a strong understanding of what that construct is and how it is exhibited. Discussions with colleagues can help us to clarify these issues. Wiggens (1998, p. 32) provides two questions that can help us make judgements about construct considerations:

- Could the student do well at the task for reasons that have little to do with the desired understanding or skill being assessed?
- Could the student do poorly at the task for reasons that have little to do with the desired understanding or skill?

Criterion considerations

Sometimes, developing a validity argument involves looking at how well our assessment results compare with or predict other measures recorded on a separate assessment or criterion. When results on two different assessments that have been designed to assess the same construct converge, we can use that as evidence that our assessments are at least "pointing in the same direction". Test developers will often carry out correlation studies that look at these relationships to help support arguments regarding validity.

In the classroom we are unlikely to carry out studies to check assessment-criterion relationships. However, we should ask questions when assessment results from different assessment tools that are meant to be testing the same construct lead us to very different interpretations of achievement. At secondary level we might carry out a study to ascertain how well our classroom assessments predict future success and whether they can be used to make valid statements about likely progress.

Consequential considerations

The final type of consideration involves evaluating the consequences of using assessment results. The weight we give to the results of an assessment will have impacts on teaching and learning. Some of these consequences can be negative, especially when the assessment format used leads to "teaching to the test" or to reduced motivation in students. For instance, when paper and pencil testing is the only form of assessment used, it can become very tempting to teach to the test and place narrow limits on classroom experiences.

We should question the validity of our assessment when there is evidence that the consequences of using the assessment results to make decisions or inform students of progress are detrimental to our overall educational goals.

Checks for validity

Our ability to make valid interpretations and decisions based on assessment data can be weakened by many factors. Being aware of these can help us frame questions that inform our decision-making about validity claims. Using some of the checks in Figure 1 can help us minimise the threats.

Final comment

Validity should be at the top of our minds when we design assessments or make decisions about assessment programmes. It is critical that our assessment results allow us to make judgements about the progress of our students that are robust and useful, and that lead to positive consequences. Being aware of validity and particularly how it can be threatened can help us make decisions about what assessments are worth making and what they can be used for.

References

Linn, R. L., & Miller, M. D. (2005). *Measurement and assessment in teaching* (9°th ed.). New Jersey: Pearson Education.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco, CA: Jossey-Bass.

Charles Darr is a senior researcher at New Zealand Council for Educational Research. He is currently working to redevelop the PAT mathematics tests and also contributes to work on the Assessment Resource Banks.

Email: charles.darr@nzcer.org.nz