# HOW TO THINK ABOUT ASSESSMENT

DYLAN WILIAM

## Introduction

In most areas of education, there are few – if any – undisputed facts. Some believe that whole language, or 'balanced literacy', is the way to teach reading in the early years, while others argue for the importance of a foundation in structured synthetic phonics. Some argue for ability grouping, particularly for older learners, and in subjects like mathematics, while others point out the advantages of mixed-ability teaching. The important point here is that in such debates, the evidence is rarely as clear as we would like, and different people can, looking at the same evidence, reasonably come to different conclusions.

Educational assessment is different. While there are areas where reasonable people can disagree, there are also many areas where you can be not just out of the mainstream, but factually, demonstrably incorrect.

The aim of this chapter is to provide the reader with an understanding of some of the most important ideas in assessment – such as reliability, validity and so on – but also to show how the right ways of thinking about these ideas can lead to much more productive discussions about how to assess. The chapter also aims to help the reader understand that there is no such thing as a perfect assessment system. Every assessment system involves trade-offs, and what matters is whether the trade-offs made are more or less appropriate in particular situations. In this way, assessment can support learning as well as measuring it.

## Assessments are procedures for drawing inferences

The word 'assessment' comes from the Latin *assidere* (originally, 'to sit with'). But then the word 'matinee' comes from the Latin *matutinus* ('early in the morning', after Matuta, the goddess of the dawn), so the origin of a word often bears no relationship to its current usage. Language evolves.

More problematically, particularly in education, we have what Truman Kelley called the 'jingle-jangle' fallacies (Kelley, 1927). The jingle fallacy (originally proposed by Herbert A. Aikins) is to assume that things with the same label are in fact the same. The jangle fallacy is the opposite: assuming that things with different labels are in fact different.

Some people use the words 'assessment' and 'evaluation' interchangeably, while others give the two terms different meanings. To make things even worse, there is not even any consistency, amongst those who distinguish between assessment and evaluation, what the difference is. For example, in higher education, some use 'assessment' to mean the process of collecting and documenting evidence for the purpose of improving learning, while 'evaluation' is used for the process of assigning grades or scores to students' performance. In compulsory education, particularly in the US, it is more common to apply the term 'assessment' to individuals, and 'evaluation' to institutions, or artifacts such as curriculum. The important point here is that there is no consensus on the meanings of the terms 'assessment' and 'evaluation', so when we use these terms, it is useful to be clear about what we mean.

In this chapter, I am going to follow Lee J. Cronbach (1971) in defining 'assessment' as a procedure for drawing inferences. We give students things to do – such as tasks, activities, tests and so on – and we collect evidence from the students, from which we draw conclusions. The conclusions may be about status, such as 'this child knows 50% of his number facts' or 'this pupil is likely to be successful in training to be a doctor', or it could be about next steps in teaching, such as 'this child is having particular difficulty with the seven times table' or 'this student seems to be having particular difficulty with electron pair repulsion theory in their mock A level Chemistry exam'.

Defining an assessment as a procedure for drawing inferences also clarifies that it makes no sense to define the terms 'formative' and 'summative' as kinds of assessment, because the same assessment can be used summatively or formatively. In the example above, a test of number facts gave us evidence for both a summative conclusion (this child knows 50% of his number facts) and a formative one (this child would likely benefit from work on the seven times table). There is no such thing as *a* formative assessment or *a* summative assessment. There are, instead, formative and summative uses of assessment information. Now to be sure, some assessments may be better for summative functions, and some may be better for formative functions; but it is the inferences, and not the assessments themselves, nor even the evidence generated by those assessments, that are formative or summative.

## Validity is a property of inferences, not of assessments

The idea that assessments are procedures for drawing inferences also helps clarify the idea of the validity of an assessment. Traditionally, 'validity' has been defined as the extent to which an assessment assesses what it purports to assess. However, there are two problems with this definition. The first is that assessments do not purport anything. The purporting (if there is such a word) is done by humans, and assessments are often used in ways that were never intended, or even envisaged, by the assessment developers. For example, GCSEs were originally intended to provide information about the extent to which a student had learned the contents of the syllabuses for their chosen subjects, ostensibly with a view to providing information about that student's potential for further study. However, GCSE grades are now used to inform judgements about the quality of education provided by the school that the student attends – a function they were never designed to fulfil (and do not do particularly well). A student's average GCSE grade may provide some information about the extent of that student's achievement on her GCSE courses, but it provides very little information about the quality of education received by that student, since the most important factors in a student's GCSE grades are nothing to do with the school, but rather the personal characteristics – and the social background – of the student (Wiliam, 2012).

The second problem with defining validity as a property of a test or other form of assessment is that an assessment can be valid in some circumstances but not others. If we had an arithmetic test with a high reading demand, what can we conclude from a student's score on the test? If the student is a fluent reader, then, provided the test samples all aspects of arithmetic, we can reasonably conclude that high scores indicate good arithmetic ability and low scores indicate weak arithmetic ability.[3] But if some of the students who take the test are weak readers, we do not know what a low score means. It might be that the student was unable to do the arithmetic, but it might mean that the student was able to do the arithmetic, but was unable to read the questions well enough to know what she was being asked to do. If we believe that validity is a property of a test, we would be in the curious situation of saying that the same test would be valid for some students but not others.

This is why there is now widespread agreement amongst assessment researchers that validity is not a property of assessments but of *inferences*. For a given

3. The word 'ability' is used here in its most literal sense, which is how able a student is to do something, and does not imply that ability is in any sense fixed. If a student learns more, then ability increases.

assessment, some conclusions will be valid, but others will not. Whether a particular assessment can support valid inferences will depend on the students to whom it is given, but it will also depend on the circumstances under which the assessment is administered.

To see why the way an assessment is administered has an impact on what kinds of inferences are supported by the results yielded by the assessment, consider a spelling test in which students are asked to spell 20 words drawn at random from a word bank of 1000 words. If a student does not know which words have been chosen, then if a student spells 10 of the 20 words correctly, then it is reasonable to assume that the student knows how to spell approximately half of the 1000 words in the word bank. But if the student knows which 20 words will be on the test, then all we know is that the student knows how to spell the 10 words they spelled correctly in the test.

This example illustrates a very important principle in assessment. When we give students an assessment, we are hardly ever interested in how well a student did on that test. We are interested in how the results on the test allow us to draw conclusions about things that were *not* on the test. The more predictable an assessment is, the less information the test provides about the things that were not tested. This does not mean that tests should not be predictable – there are times when this is entirely appropriate – but it is important to realize that a predictable test tells us only about the things that were tested. The results a student gets on a test on adding three-digit numbers in which all the sums are in vertical format:

$$\begin{array}{rrrr} & 5 & 6 & 4 \\ + & 3 & 6 & 7 \\ \hline & & & \\ \hline \end{array}$$

will not tell us whether a student can do the same calculation in horizontal format:

**564 + 367 =**

This is why validity cannot be a property of a test. A test is valid for some conclusions, but not others. This was nicely summed up by Cronbach (1971): 'One validates, not a test, but an *interpretation of data arising from a specified procedure*.' (p. 447, emphasis in original)

So, when someone asks, 'Is this test valid?' in my view the best response is, 'Tell me what you propose to conclude about a student when you see their test score, and I'll tell you whether that conclusion is justified.'

The focus on inferences, rather than assessments, also helps clarify the issue of bias in assessment. Regarding bias as a property of assessments runs into the same problems as regarding validity as a property of tests. After all, a test tests what a test tests. The bias comes when we conclude that a particular assessment outcome has a particular meaning. Bias, like validity, is a property of inferences, not of assessments.

## There are two main threats to validity

There are two main reasons why the results from assessments may not support the conclusions we want to make. The first is that the assessment does not assess the sorts of things which we want to make inferences about – intuitively, the assessment is 'too small'. The second is that the assessment assesses things which are not relevant to the things which we want to make inferences about – the assessment is, in some sense, 'too big'.

### Construct underrepresentation

The technical term for the first reason (when the assessment is too small) is 'construct underrepresentation'. The idea here is that we have a construct of interest – say science achievement – and the assessment does not cover all the things that we would need to know about a student to draw conclusions about the student's achievement on this construct.

Of course, whether this is an issue depends on how we define our construct of interest, but if we define science achievement so as to include practical skills, then our assessment must include practical assessments or else we cannot be sure that a student's performance on a written assessment is a good guide to their practical skills.

In response to this, people often counter by saying that the scores that students get on practical assessments correlate highly with their scores on written tests, so it is really a waste of money to include practical assessments; we can use the scores on the written assessment as a proxy for the scores on the practical assessment. While this may be true as long as teachers are including practical work in their curriculum, failing to assess all important parts of a subject makes it possible to increase a student's score on a test by ignoring the untested parts of the curriculum. When schools are under pressure to increase test scores, narrowing the curriculum to focus only on the things that are tested makes it easier to increase students' achievement on the things that are measured. This is a bit like putting ice cubes in the mouth of a patient with a fever. When you measure the patient's temperature with a thermometer in the mouth, you get a

lower reading on the thermometer, but you haven't addressed the underlying issue, which is the fever. You have changed the indicator, but not the indicated.

Now it is important to realize that if a test or other form of assessment assesses all the important aspects of a subject but is used in a high-stakes setting, and teachers teach to the test, then that does not jeopardize the validity of the assessment. The assessment assesses everything it should. But if the assessment underrepresents the subject, such as a test of English language that does not assess speaking and listening, and in a high-stakes setting, teachers reduce the amount of time they spend on the untested aspects, then the validity of the assessment is in question, because it was the deficiencies in the assessment (the construct underrepresentation) that caused the adverse social consequences.

## Construct-irrelevant variance

The technical term for the second issue is 'construct-irrelevant variance', which seems like the worst kind of psychological jargon, but it is worth taking time to understand this idea because it can help us think about assessment problems in more powerful and productive ways.

Recall the arithmetic test with a high reading demand discussed earlier. Ideally, with an arithmetic test, we would want differences in scores on the test to be associated with differences in arithmetic ability, and only differences in arithmetic ability. If all students taking the test are fluent readers, and the test is a good test of arithmetic, then variations in the scores achieved will be due to differences in arithmetic ability. But if some of the students are weak readers and the reading demand of the test is high, some of the variation in the scores on the test will be caused by differences in arithmetic ability and some by differences in reading ability. Variation in scores caused by variation in arithmetic ability is construct-relevant – after all, this is what we are trying to assess. But variation in scores caused by variation in reading skill is construct-irrelevant – differences in reading ability should not affect a student's score on an arithmetic test. And because statisticians tend to measure variation in scores by calculating the variance of a set of scores,[4] when scores are influenced

4. To find the variance of a set of scores, we find the mean of all the scores and subtract each score from the mean. It makes no sense to average the resulting numbers because the mean will be zero, so first, we square each of these differences from the mean (thus getting rid of all minus signs) and find the average of the resulting numbers. This is the variance, and is a measure of how spread out the scores are.

by things that should not be influencing the scores, the scores are said to suffer from construct-irrelevant variance.

From the foregoing, it should be clear that construct-irrelevant variance is a property of a set of scores, not of the assessment itself. If we gave our arithmetic test to fluent readers, the variation in the scores would be construct-relevant (because the reading proficiency of the students taking the test would not be an issue). However, if some students taking the test are poor readers, then some of the variation in scores will be caused by differences in reading ability, so there would be a degree of construct-irrelevant variance in the scores. The variation in scores on a particular test might be construct-relevant for one group of students, but include some construct-irrelevant variation with another group of students.

## The importance of construct definition

To see how these two threats to validity – construct underrepresentation and construct-irrelevant variance – can help clarify our thinking, it is useful to consider how we might assess a student's knowledge of history, and in particular, whether we can assess our students' knowledge of history just by using multiple-choice questions.

Some people say yes, and some people say no, and this debate appears to be a debate about the suitability of different methods of assessment; but in reality, people on different sides of this argument have different beliefs about what it means to be good at history – i.e., the construct of history.

For those who think history is all about facts and dates, multiple-choice questions are pretty nifty because you can assess a lot of facts and dates in a reasonably short period of time. Moreover, the 'facts-and-dates' brigade regard essay questions as inappropriate because while some of the variation in scores on such questions will be due to differences in historical knowledge, some of it will be due to differences in writing ability, and even handwriting speed. In other words, the 'facts-and-dates' brigade will regard scores on history assessments that involve essay writing as embodying some construct-irrelevant variance (those better at writing do better).

On the other hand, those who think that being good at history is more than just knowing facts and dates and also includes things like being able construct historical arguments will regard assessments made up entirely of multiple-choice questions as underrepresenting the construct of history. Students who

are better at constructing historical arguments will do no better than those who are not if they do equally well on facts and dates.

The important point here is that the debate about how to assess history appears to be a debate about assessment methods, but in reality it goes much deeper. The debate does come to surface when we are talking about how to assess history, but in reality it is an argument about what it means to be good at history – it is an argument about how the construct of historical knowledge should be defined.

This matters, because if the construct has been defined properly, different people should agree about whether a particular set of assessments adequately samples the domain of interest. In other words, with good construct definition, assessment design is a largely technical matter. However, if the construct is not well-defined, then assessment design becomes a value-laden process. In particular, the values of the people designing the assessment play a part in the design of the assessment.

Now that we have tools for thinking about validity, it might seem obvious that we should turn our attention to the other desirable quality of assessments: reliability. However, we do not need to do so because we already have the tools we need, because reliability is actually part of validity.

## (Un)reliability is an aspect of construct-irrelevant variance

As mentioned earlier, in the example of the arithmetic test with a high reading demand, when the test is given to both weak and strong readers, there is an element of construct-irrelevant variance in the scores. This construct-irrelevant variance is systemic in that it is likely to affect all poor readers in a similar way. However, some sources of construct-irrelevant variance are random. Students have good days and bad days, so the performance in a test on a particular occasion might not be typical of what that student would achieve on other occasions. One marker might give a student the benefit of the doubt on a particular question while another would not. The particular questions included in a test might suit some students better than others.

Traditionally, factors such as these have been regarded as issues of reliability, and people have talked about the need for assessments to be 'valid and reliable', implying that validity and reliability were separate properties of assessments. However, such a perspective makes little sense because if the results of an assessment are unreliable, then they cannot possibly support valid inferences. If the score a student gets tomorrow is very different from the score they got today,

then any conclusions that you draw about that student's capabilities on the basis of today's test score are unlikely to be valid. Reliability is a pre-requisite for validity.

However, there seems to be a paradox here, because reliability is often seen to be in tension with validity, with attempts to increase reliability having a negative influence on validity. How can reliability be in tension with validity while at the same time be a pre-requisite for it?

The two threats to validity discussed above – construct underrepresentation and construct-irrelevant variance – provide a resolution of this paradox. Increasing reliability – for example by standardizing assessments, by giving markers strict marking guidelines, and focusing only on aspects of a subject that are easy to assess – may well reduce construct-irrelevant variance, but only at the expense of reducing the representation of the construct (or, in other words, *increasing* the amount of construct underrepresentation). Put simply, we are prioritizing the reduction of some threats to validity at the expense of others. For a given amount of assessment time, we can cast our net widely and get some not particularly reliable information about a large number of aspects of a subject, or we can focus our attention on much more limited aspects of a subject and get much more reliable information. And of course, there is no right answer here. Sometimes we need a floodlight to get a perspective on a wide area, and sometimes we need a spotlight, getting clear information about a small area. What matters is whether the trade-off between reliability and other aspects of validity is more or less appropriate for the particular situation.

This last point is particularly important because it is often assumed that more reliability is better, but unless we narrow the assessment (and therefore increase construct underrepresentation) the only way to make an assessment more reliable is to make it longer. Moreover, the increases in testing time needed to make assessments more reliable are substantial. For example, to reduce the number of students getting the wrong grade in a GCSE subject from 40% to 25% would require increasing the length of the exams in that subject fourfold.[5]

Because appreciable increases in the reliability of our assessments require so much extra assessment time – time that would be better spent on teaching our students – we need to understand how reliable our assessments are, so that

5. The estimate is based on applying the Spearman-Brown prophecy formula to the data in the table on page 19 of Wiliam (2001).

we can make informed judgements about how much weight to place on the information they provide.

## Reliability isn't everything, but it is important

The starting point for estimating the reliability of an assessment is to assume that a student has a true score on that assessment. When people hear the term 'true score', they assume that this means assuming that ability is fixed, but this is not the case. The true score is simply the long run average that a student would score over many administrations of a test, assuming that no learning takes place.

For example, returning to the 20-word spelling test discussed earlier, if the student actually knows how to spell 600 of the 1000 words in the word bank, her true score is 60%. If she learns how to spell another 100 words her true score will be 70%. If she forgets how to spell 100 of these words, her true score will be 50%. We could of course find a student's true score by testing her on all 1000 words in the bank, but we have better things to do with our (and her) time, so we take a random sample of all the things we might assess and use the student's score on the sample as an estimate of her score on the whole domain (in this case, the 1000 words in the word bank). The reliability of our test is simply an indication of how good the score on the test is as a guide to the student's proficiency on the whole word bank.

Suppose we ask a student to spell 20 of the 1000 words in the word bank, drawn at random, on five occasions over the course of the day, and her scores are 15, 17, 14, 15 and 14. On average, she scores 15 out of 20 (i.e., 75%) so our best guess is that she can spell 750 of the 1000 words.

If, on the other hand, the results had been 20, 12, 17, 10 and 16, her average score would still be 15 out of 20, so our best guess would still be that she knows 750 of the 1000 words, but now we would be much less confident that our samples were a good guide to the whole bank because the scores vary so much. An analogy might be helpful here. If you wanted to find out if your child had a fever, and had only an electronic thermometer available, then if you measured the child's temperature three times, and got 36.6°C, 37.4°C, and 37.0°C you would probably feel confident that your child did not have a fever (conventionally defined as a body temperature over 38.0°C). But if the readings you got were 35.0°C, 39.0°C, and 37.0°C, you might well be alarmed. In both cases, the average of the three readings is the same (37°C) but in the second case, the spread of the readings gives us little confidence in the readings.

To create a measure of how spread out our scores are, we subtract the average score from each score, so in the first example, we would get the following:

| | | | | | |
|---|---|---|---|---|---|
| **Actual score** | **15** | **17** | **14** | **15** | **14** |
| **Difference from average** | **0** | **+2** | **−1** | **0** | **−1** |

If we took the average of these five differences, we would get zero (that is, after all, the definition of the average), so we square each of the differences (to get rid of the minus signs), add up the totals, and then divide by the number of scores (in this case, five). We then take the square root of the result, which gives us the standard deviation of the errors, which in this case is 1.2.

Using the scores in the second example, we would get

| | | | | | |
|---|---|---|---|---|---|
| **Actual score** | **20** | **12** | **17** | **10** | **16** |
| **Difference from average** | **+5** | **−3** | **+2** | **−5** | **+1** |

The standard deviation of the errors here is 4.0, which tells us that the score that a student gets in this case would be a less accurate guide to the score a student might get on the whole test.

With a typical school test or other form of assessment, the errors will usually form a classic 'bell curve' or normal distribution, so we can use the properties of the normal distribution to see what these numbers mean. With a normal distribution, 68% of the data points fall within one standard deviation of the mean and 96% fall within two standard deviations of the mean. Therefore, if the standard deviation of the errors – often called the standard error of measurement or SEM – for all students was 1.2, then for approximately two-thirds of the students in a group, their score on any one assessment occasion will be within 1.2 points of their true score, and almost all (96%, or 24 out of 25) will get a score within 2.4 points of their true score.

If, however, the SEM is 4 points, then for two-thirds of the students, their actual score will be within 4 points of the true score, and for 96% of the students, their score will be within 8 points of the true score. However, for every class of 25, there will be one student who gets a score that is more than 8 points away from their true score. Unfortunately, we won't know which student this is, nor whether the score they got was too high or too low.

These of course are just examples. To find out what this looks like in practice, it is useful to see the errors of measurement that are typical with educational assessments. Most test publishers do report the standard errors of measurement of their assessments, but tend to give more prominence to a test's index of reliability. This ranges from 0 to 1, with a completely random assessment (i.e., one where a student's score is completely a matter of chance) having a reliability of 0 and a perfectly reliable assessment (where the student would get exactly the same score on every testing occasion) having a reliability of 1. To see how the SEM relates to the reliability, the table below shows the relationship for an assessment where the average score for all students is 50% and where almost all students score between 20% and 80% (i.e., a standard deviation of 15).[6]

| **For a typical test (average score 50, standard deviation 15), a student with a true score of 60 will, on a given occasion, score** | | | |
|---|---|---|---|
| Reliability | SEM | Two-thirds of the time (68%) | Almost always (96%) |
| 0.70 | 8.2 | 52 to 68 | 44 to 76 |
| 0.75 | 7.5 | 53 to 68 | 45 to 75 |
| 0.80 | 6.7 | 53 to 67 | 47 to 73 |
| 0.85 | 5.8 | 54 to 66 | 48 to 72 |
| 0.90 | 4.7 | 55 to 65 | 51 to 69 |
| 0.95 | 3.4 | 57 to 63 | 53 to 67 |

Teacher-produced tests typically have reliability indices in the range 0.70 to 0.80, while standardized tests have reliability from 0.90 to 0.95, and specialized psychological tests can have reliabilities over 0.95. The reliability of GCSE exams ranges from 0.70 to 0.95 with a mean value around 0.83 (Hayes and Pritchard, 2013)

Reliability data for the 2018 key stage 2 tests are shown in the table below (Thomson, 2019).

6. Formally, the relationship between the reliability index, *r*, and the standard error of measurement (SEM) is given by the formula SEM = SD × $\sqrt{(1 - r)}$ where SD is the standard deviation of the scores of all the students taking the test. When *r* is zero, the SEM is equal to the SD of all the scores, because the test is providing no information about the student. Our uncertainty about a student's score is just as great after we are told the result (the SEM) as before (the whole group SD). When *r* is 1, then the SEM is zero, because there is no uncertainty about the student's result.

| Test | Duration | Reliability | SEM (%) |
|---|---|---|---|
| Mathematics | 110 minutes | 0.96 | 5.3 |
| Reading | 60 minutes | 0.90 | 5.9 |
| Grammar, punctuation, spelling | 60 minutes | 0.95 | 4.5 |

In other words, in 2018, two-thirds (actually 68%) of students received a score within 5.9% of their true score in reading, within 5.3% in mathematics and within 4.5% in grammar, punctuation and spelling (GPS). However, in a class of 25 students, one student would have got a score at least 11% lower or higher than their true score in mathematics, at least 12% different from their true score in reading, and 9% off in GPS.

As noted above, this does not necessarily mean that we want more reliable tests. For example, to make the reading test as reliable as the maths test, we would need to increase the length of the test to 180 minutes, which might well bring additional problems such as student fatigue. Instead, the important message here is that the reliability of our tests may well be optimal. We do not necessarily need more reliable tests. What we do need is to be aware of the limitations of our assessments so that we do not place more weight on the result of an assessment than its reliability would warrant.

This is particularly important when looking at change scores – the change in a student's score over a period of time – because we are, in effect, subtracting one unreliable number from another unreliable number. Indeed, the unreliability of change scores has led some psychologists to conclude that we should not even try to measure change (Cronbach and Furby, 1970). The problem is that, in education, change is what we are mostly interested in. We have to measure change, because it is the most important thing in education, but we need to be cautious about the meanings of these change scores. For example, it is common to find that the standard error of measurement of, say, a standardized reading test is approximately the same as the progress that an average student makes in six months. In other words, we end up saying, 'Over the last six months, you have made six months' progress, give or take six months.'

Finally, it is worth looking to see how reliability and other aspects of validity interact when we use assessments to make decisions about how to group students. The table below shows how accurate our placement of students into four ability groups or 'sets' would be if we assign students to sets on the basis of a test with reliability of 0.9 and where the correlation of the score on the test

with eventual mathematics achievement (what is sometimes called 'predictive validity') is 0.7 (both of these figures are about as good as we can expect).[7]

| | | …should be in… | | | |
|---|---|---|---|---|---|
| | | set 1 | set 2 | set 3 | set 4 |
| Students actually placed in… | set 1 | 23 | 9 | 3 | |
| | set 2 | 9 | 12 | 6 | 3 |
| | set 3 | 3 | 6 | 7 | 4 |
| | set 4 | | 3 | 4 | 8 |

In other words, by looking at the numbers that are not in the leading diagonal of the table above, we can see that 50 of the students are in the 'wrong' set.

This is not to make an argument for or against ability group in schools – the research on this issue is nowhere near as clear-cut as some people claim, so judgment is required. It is to point out that even if our assessments are as good as the 'state of the art', they are far from perfect, and we need to be cautious in making decisions on the basis of test results, especially when these decisions have profound consequences for students.

## Summary

The key conclusions of this chapter are:

**There is no such thing as a valid test.** Rather, validity is best thought of as a property of inferences based on test outcomes. An assessment will, depending how it is administered, support some inferences, but not others. Moreover, a test may support valid inferences for some students and not others. In a similar vein, there is no such thing as a formative assessment or a summative assessment, because formative and summative are properties of inferences, not of assessments.

**There are two main threats to validity: construct underrepresentation and construct-irrelevant variance.** Some assessments are, in a sense, too small. They do not provide us with information about things we need to know about to draw the conclusions we want to draw, so they underrepresent the construct of interest. Some assessments, on the other hand, are too large. They may assess

7. I have also assumed that, as is typical, the higher-achieving sets are larger than those for lower-achieving students.

the thing we want to know about but students' results are also affected by things that are unrelated to the things we want to know about, so the scores students get vary for reasons that are irrelevant to the things we want to know about (hence construct-irrelevant variance).

**Arguments about assessment methods are often (usually?) arguments about constructs.** When people find it hard to agree on whether a particular assessment method is appropriate, it is often, and perhaps usually, because they disagree about what should be assessed.

**(Un)reliability is the random component of construct-irrelevant variance.** When student performance varies from occasion to occasion, when the same work is given different marks by different markers (or even the same marker on different occasions), when the particular selection of questions included in the assessment – when any of these influence a student's score, there is random variation in the scores that is irrelevant to the construct of interest.

**Change scores are much less reliable than status scores.** While we do need to know about change scores – after all, we want to know whether our students are getting better – we need to be especially cautious in interpreting change scores, because they are the result of subtracting one unreliable number from another.

**More reliability is not necessarily better.** Assessments have to be made much longer to have a significant impact on reliability, taking time away from teaching. Relatively low reliability may be optimal, provided we know how reliable an assessment is, and therefore, how much weight to place on it.

**All assessment involves trade-offs.** The most important concept in education is opportunity cost: time that you spend assessing your students is time that you (and they) do not have for other things. The key thing in assessment is being clear about why you are assessing, what conclusions you want to draw, and how well your evidence supports the conclusions you want to draw. Keep those three things in mind, and you won't go far wrong.

**References**

Cronbach, L. J. (1971) 'Test validation' in Thorndike, R. L. (ed.) *Educational measurement*. 2nd edn. Washington, DC: American Council on Education, pp. 443–507.

Cronbach, L. J. and Furby, L. (1970) 'How we should measure "change" – or should we?', *Psychological Bulletin* 74 (1) pp. 68–80.

Hayes, M. and Pritchard, J. (2013) *Estimation of internal reliability*. Coventry: Ofqual.

Kelley, T. L. (1927) *Interpretation of educational measurements*. Yonkers-on-Hudson, NY: World Book Company.

Thomson, D. (2019) 'How reliable are key stage 2 tests?', *FFT Education Datalab* [Blog], 3 April. Retrieved from www.bit.ly/2yLKdjZ

Wiliam, D. (2001) 'Reliability, validity, and all that jazz', *Education 3–13* 29 (3) pp. 17–21.

Wiliam, D. (2012) 'Are there "good" schools and "bad" schools?' in Adey, P. S. and Dillon, J. (eds) *Bad education: debunking myths in education*. Maidenhead: Open University Press, pp. 3–15.

## Author bio-sketch:

Dylan Wiliam is emeritus professor of educational assessment at UCL. In a varied career, he has taught in inner-city schools, trained teachers, directed a large-scale testing program, and served a number of roles in university administration, including dean of a school of education. His research focuses on supporting teachers to develop their use of assessment in support of learning.