Reliability, validity, and all that jazz

Dylan Wiliam

King's College London

To appear in Education 3-13 ***29****(3) 9-13 (2000)*

Introduction

No measuring instrument is perfect. If we use a thermometer to measure the temperature of a liquid, even if the temperature doesn't change, we get small variations. We generally assume that these are random fluctuations around the true temperature, and we take the average of the readings as the temperature of the liquid but there are more subtle problems. If we place the thermometer in a long narrow tube of the liquid, then the thermometer itself might affect the reading—if the thermometer is warmer than the liquid, the thermometer will heat up the liquid. In contrast to the problems of reliability mentioned above, this effect is not random, but a systematic bias, and so averaging across a lot of readings does not help. Sometimes, we are not really sure what we are measuring. For example bathroom scales sometimes change their reading according to how far apart the feet are placed. The scales are not just measuring weight but also the way the weight is distributed. While random fluctuations in a measurement affect the *reliability* of a measurement, the problem of bias and the related problem of being clear about what exactly we are measuring are aspects of the *validity* of the measurement.

These ideas extend in natural ways to educational assessments and the purpose of this article is, through the use of large-scale simulations, to illustrate the fundamental limitations of this process—to spell out what educational assessments can, and more importantly, cannot, do for us.

Reliability

If a student attempts a test several times, even if no learning takes place, the student will not get the same score each time—the student might not feel very 'sharp', the marker may be more or less generous, or the handwriting might be a little bit clearer so the marker can understand the answer. A further source of unreliability (usually the largest) concerns the particular choice of items. A test is constructed by choosing a set of items from a much bigger pool of potential items. Any particular set of items that are actually included will benefit some students (eg those that happen to have revised those topics recently) and not others.

These fluctuations affect the quality of the information that a test gives us. For a good test, the size of fluctuations must be small in comparison with the amount of information about the individual—drawing a parallel with communications engineering, we want a good signal-to-noise ratio.

The starting point for estimating the reliability of a test is to hypothesise that each student has a 'true score' on a particular test—this does not mean that we believe that a student has a true ability at, say, reading, nor that the reading score is in any sense fixed. An individual's true score on a test is simply the average score that the individual would get over repeated takings of the same or a very similar test. We can then say that a student's actual score on any particular occasion (usually denoted X) is made up of their true score, T (ie what they 'should' have got) plus a certain amount of 'error', E (ie a measure of the extent to which the result on particular testing occasion departed from the true score). We can therefore write:

$$X = T + E$$

On a given day, a student might get a higher score than their true score (in which case E would be positive), or a lower score (in which case E would be negative). In order to get a measure of reliability, we need to be able to compare the sizes of the errors (E) with the sizes of the actual scores (X). When the errors are small in comparison with the actual scores, we have a relatively reliable test, and when the errors are large in comparison with the actual scores, then we have a relatively unreliable test.

We cannot, however, use the average values for this comparison, because, by definition, the average value of the errors is zero! Instead, we use a measure of how spread out the values are, called the standard deviation. The key formula is

$$\text{standard deviation of errors} = \sqrt{1-r} \times \text{standard deviation of observed scores}$$

where r is the reliability coefficient (or just the reliability) of the test.

A coefficient of 1 means that the standard deviation of the errors is zero, so there is no error, so the test is perfectly reliable. A coefficient of 0 means that the standard deviation of the errors is the same as that of the observed scores—the scores obtained by the individuals are all error, so there is no information about the individuals at all! When a test has a reliability of zero the result of the test is completely random.

The reliability of tests produced in schools is typically around 0.7 to 0.8, while that for commercially produced educational tests range from 0.8 to 0.9 and can be over 0.9 for specialist psychological tests. To see what this means in practice, it is useful to look at some specific kinds of tests.

The reliability of standardised tests

Knowing a student's mark on a test is not very informative unless we know how hard the test is. Because calibrating the difficulty of tests is complex, the results of many standardised tests are reported on a standard scale which allows the performance of individuals to be compared with the performance of a representative group of students who took the test at some point in the past. When this is done, it is conventional to scale the scores so that the average score is 100 and the standard deviation of the scores is 15. This means that

- 68% (ie roughly two-thirds) of the population score between 85 and 115
- 96% score between 70 and 130

So we can say that the level of performance of someone who scores 115 on a reading test would be achieved by 16% (ie half of 32% which is 100% - 68%) of the population, or that this level of performance is at the 84th percentile.

From this it would be tempting to conclude that someone who scored 115 on the test really is in the top 16% of the population, but this may not be the case, because of the unreliability of the test. If we assume the test has a reliability of 0.85 (a reputable standardised test will provide details of the reliability, and how it was calculated), then we can estimate the likely error in this score of 115.

Since the standard deviation of the scores is 15, and reliability is 0.85, from our key formula we can say that the standard deviation of the errors is

$$\sqrt{1-0.85} \times 15$$

which is just under 6.

The standard deviation (SD) of the errors (often called the standard error of measurement or SEM) tells us how spread out the errors will be on this test:

- For 68% of the candidates their actual scores will be within 6 (ie one SD) of their true scores
- For 96% of the candidates their actual scores will be within 12 (ie two SDs) of their true scores
- For 4% of the candidates their actual scores will be at least 12 away from their true score.

For most students in a class, their actual score will be close to their true score (ie what they 'should' have got), but for at least one student, the score is likely to be 'wrong' by 12 points (but of course we don't know who this student is, nor whether the score they got was higher or lower than their true score). For a test with a reliability of 0.75, the SEM is 7.5, so someone who scores 115 (who we might think is in the top sixth of the population) might really have a true score of just 100 making them average or as high as 130, putting them in the top 2% (often used as the threshold for considering a student 'gifted').

Because the effects of unreliability operate randomly, the averages across *groups* of students, however, are quite accurate. For every student whose actual score is lower than their true score, there is likely to be one whose actual score is higher than their true score, so the average observed score across a class of students will be the same as the average true score. However, just as the person with one foot in boiling water and one foot in ice was quite comfortable 'on average' we must be aware that the results of even the best tests can be wildly inaccurate for individual students, and therefore high-stakes decisions should never be based on the results of individual tests.

National curriculum tests

Making sense of reliability for national curriculum tests is harder because we use levels rather than marks, for good reason. It is tempting to regard someone who gets 75% in a test as being better than someone who gets 74%, even though the second person actually might actually have a higher true score. In order to avoid unwarranted precision, therefore, we often just report levels. The danger, however, is that in avoiding unwarranted precision, we end up falling victim to unwarranted accuracy—while we can see that a mark of 75% is only a little better than 74%, it is tempting to conclude that level 2 is somehow qualitatively better than level 1. Firstly, the difference in performance between someone who scored level 2 and someone who scored level 1 might be only a single mark, and secondly, because of the unreliability of the test, the person scoring level 1 might actually have a higher true score.

Only limited data have been published about the reliability of national curriculum tests, although it is likely that the reliability of national curriculum tests is around 0.80—perhaps slightly higher for mathematics and science . Nevertheless, by creating simulations of a 1000 students at a time, we can see how the proportion of students who would be awarded the 'wrong' levels at each key stage of the national curriculum varies as a result of the unreliability of the tests as shown in table 1.

reliability of test	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
% of students misclassified at KS1	27%	25%	23%	21%	19%	17%	14%	10%
% of students misclassified at KS2	44%	42%	40%	36%	32%	27%	23%	16%
% of students misclassified at KS3	55%	53%	50%	46%	43%	38%	32%	24%

Table 1: variation in proportion of misclassifications in national curriculum tests with reliability

It is clear that the greater the precision (ie the more different levels into which we wish to classify people), the lower the accuracy. What is also clear is that although the proportion of mis-classifications declines steadily as the reliability of a test increases, the improvement is very slow.

Making tests more reliable

We can make tests more reliable by improving the items included in the tests, and by making the marking more consistent, but in general, the effect of such changes is small. There are only two effective ways of increasing the reliability of a test: make the scope of the test narrower, so you ask more questions on the same topic, or, what amounts to the same thing, make the test longer. In general if we have a test of reliability r and we want a reliability of R, then we need to lengthen the test by a factor of n given by

$$n = \frac{R(1-r)}{r(1-R)}$$

So, if we have a test with a reliability of 0.75, and we want to make it into a test with a reliability of 0.85 we would need a test 1.9 times as long. In other words, doubling the length of the test would reduce the proportion of students mis-classified by only 4% at key stage 1, by 9% at key stage 2 and by 6% at key stage 3. It is clear that increasing the reliability of the test has only a small effect on the accuracy of the levels. In fact, if we wanted to improve the reliability of key stage 2 tests so that only 10% of students were awarded the incorrect level, we should need to increase the length of the tests in each subject to over 30 hours[1].

Now it seems unlikely that even the most radical proponents of schools tests would countenance 30 hours of testing for each subject. Fortunately, there is another way of increasing the effective length of a test, without increasing testing time, and that is through the use of teacher assessment. The experience of GCSE has shown that the danger of bias in teacher assessments can be adequately addressed through standardisation and moderation. By using teacher assessment, we would in effect, be using assessments conducted over tens, if not hundreds of hours for each student, producing a degree of reliability that has never been achieved in any system of timed written examinations.

Using tests to predict future performance

As well as certifying achievement, one of the most common uses of tests is to predict future performance—the usefulness of a test for this purpose depends entirely on the correlation between the

scores on the test (usually called the predictor) and the scores on whatever we are trying to predict (usually called the criterion).

For example, we might, like most secondary schools in the UK, want to use the results of IQ tests taken at the age of 11 to predict scores on GCSE examinations taken at 16. What we would need to do would be to compare the GCSE scores obtained by students at age 16 with the scores the *same* students obtained on the IQ tests five years earlier, when they were 11. In general we would find that those who got high scores in the IQ tests at 11 get high grades in GCSE, and low scorers get lower grades. However, there will also be some students getting high scores on the IQ tests that do not go on to do well at GCSE and vice-versa. How good the prediction is—often called the predictive validity of the test—is usually expressed as a correlation coefficient. A correlation of 1 means the correlation is perfect, while a correlation of zero would mean that the predictor tells us nothing at all about the criterion. Generally, in educational testing, a correlation of 0.7 between predictor and criterion is regarded as good.

However, in interpreting these coefficients, care is often needed because such coefficients are often reported after 'correction for unreliability'. The validity of IQ scores as predictors of GCSE is usually taken to mean the correlation between true scores the predictor and true scores on the criterion. However, as we have seen, we never know the true scores—all we have are the observed scores, and these are affected by the unreliability of the tests. When someone reports a validity coefficient as being corrected for unreliability, they are quoting the correlation between the true scores on the predictor and criterion, by applying a statistical adjustment to the correlation between the observed scores, which will appear to be much better than we can actually do in practice, because of the effects of unreliability. For example, if the correlation between the true scores on a predictor and a criterion—ie the validity 'corrected for unreliability'—is 0.7, but each of these is measured with tests of reliability 0.9, the correlation between the actual values on the predictor and the criterion will be less than 0.6.

Using tests to select individuals

As well as being used to predict future performance, tests are frequently used to select individuals. If we use a test to group a cohort of 100 students into 4 sets for mathematics, with, say, 35 in the top set, 30 in set 2, 20 in set 3 and 15 in set 4, how accurate will our setting be?

If we assume that our selection test has a predictive validity of 0.7 and a reliability of 0.9, then of the 35 students that we place in the top set, only 23 should actually there—the other 12 should be in sets 2 or 3, but perhaps more importantly, given the rationale given for setting, 12 students who should be in set 1 will actually be placed in set 2 or even set 3. Only 12 of the 30 students in set 2 will be correctly placed there—9 should have been in set 1 and 9 should have been in sets 3 and 4. The complete situation is shown in table 2.

In other words, because of the limitations in the reliability and validity of the test, then only half the students are placed where they 'should' be. Again, it is worth noting that these are not weaknesses in the quality of the tests but fundamental limitations of what tests can do—if anything, the assumptions made here are rather conservative—reliabilities of 0.9 and predictive validities of 0.7 are at the limit of what we can achieve with current methods. As with national curriculum testing, the key to improved reliability lies with increased use of teacher assessment, standardised and moderated to minimise the potential for bias.

		should actually be in			
		set 1	set 2	set 3	set 4
students	set 1	23	9	3	
placed	set 2	9	12	6	3
in	set 3	3	6	7	4
	set 4		3	4	8

Table 2: accuracy of setting with a test of validity of 0.7

The relationship between reliability and validity

It is sometimes said that validity is more important than reliability. In one sense this is true, since there is no point in measuring something reliably unless one knows what one is measuring. After all, that would be like saying "I've measured something, and I know I'm doing it right, because I get the same reading consistently, although I don't know what I'm measuring". On the other hand, reliability is a pre-requisite for validity—no assessment can have any validity at all if the mark a student gets varies radically from

occasion to occasion, or depends on who does the marking. To confuse matters even further, it is often the case that reliability and validity are in tension, with attempts to increase reliability (eg by making the marking scheme stricter) having a negative effect on validity (eg because students with good answers not foreseen in the mark scheme cannot be given high marks).

Of course reliability and validity are not absolutes but degrees, and the relationship between the two can be clarified through the metaphor of stage lighting. For a given amount of lighting power (cf testing time), one can use a spotlight to illuminate a small part of the stage very brightly, so that one gets a very clear picture of what is happening in the illuminated area (high reliability), but one has no idea what is going on elsewhere, and the people in darkness can get up to all kinds of things, knowing that they won't be seen (not teaching parts of the curriculum not tested). Alternatively, one can use a floodlight to illuminate the whole stage, so that we can get some idea what is going on across the whole stage (high validity), but no clear detail anywhere (low reliability). The validity-reliability relationship is thus one of focus. For a given amount of testing time, one can get a little information across a broad range of topics, as is the case with national curriculum tests, although the trade-off here is that the scores for individuals are relatively unreliable. One could get more reliable tests by testing only a small part of the curriculum, but then that would provide an incentive for schools to improve their test results by teaching only those parts of the curriculum actually tested. For more on the social consequences of assessment, see Wiliam (1992, 1996).

Summary

In this article my purpose has not been to indicate what kinds of things can and can't be assessed appropriately with tests. Rather, I have tried to illuminate how the key ideas of reliability and validity are used by test developers and what this means in practice—not least in terms of the decisions that are made about individual students on the basis of their test results. As I have stressed throughout this article, these limitations are not the fault of test developers, However inconvenient these limitations are for proponents of school testing, they are inherent in the nature of tests of academic achievement, and are as real as rocks. All users of the results of educational tests must understand what a limited technology this is.

References

Wiliam, D. (1992). Some technical issues in assessment: a user's guide. *British Journal for Curriculum and Assessment,* **2**(3), 11-20.

Wiliam, D. (1996). National curriculum assessments and programmes of study: validity and impact. *British Educational Research Journal,* **22**(1), 129-141.

---

[1] The classification consistency increases broadly as the fourth root of the test length, so a doubling in classification consistency requires increasing the test length 16 times.