Reliability, validity, and all that jazz

Dylan Wiliam

King's College London

Introduction

No measuring instrument is perfect. The most obvious problems relate to reliability. If we use a thermometer to measure the temperature of a liquid, even if the temperature of the liquid remains constant, we usually get small variations in the readings. We generally assume that these variations are random fluctuations around the true temperature of the liquid, and we take the average of the readings we got as the temperature of the liquid. However, there are also more subtle problems with measurements. If we take a thermometer and place it in a long narrow tube of the liquid, then the thermometer itself might affect the reading. For example, if the thermometer is warmer than the liquid, the thermometer could heat up the liquid, changing the result. In contrast to the problems of reliability mentioned above, these effects are not random, but a systematic bias, and so averaging across a lot of readings does not help. In other cases, we are not really sure what we are measuring. For example, some bathroom scales, because of their design, change their readings according to how far apart the feet are placed. Rather than just measuring weight, therefore, the scales are measuring something that depends on weight, but also depends on the way that the weight is distributed. While the random fluctuations in a measurement are generally regarded as affecting the *reliability* of a measurement, the problem of bias and the related problem of being clear about what, exactly, we are in fact measuring, are generally regarded as aspects of the *validity* of the measurement.

These ideas extend in fairly natural ways to educational assessments. Although in the United Kingdom we tend to resist the term 'Educational Measurement', which is common in the United States, whenever we put a mark, grade, score, or label of any kind on a piece of work done by a student, we are, in fact, making a measurement. 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 perhaps more importantly, cannot, do for us.

Reliability

If a student attempts a test several times, even if we assume that no learning takes place between administrations of the test, the student will not gain exactly the same score each time—the student might not feel very 'sharp' on one occasion or the marker may be a little more generous on one occasion than another. On another occasion, a student's handwriting might be a little bit clearer so the marker can make sense of the method being used. Although these are the most generally cited sources of unreliability, there is a third source, generally much more significant than the first two, which is caused by 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.

The purpose of any test is to give us information about an individual. As we have seen above, any measurement entails a certain amount of error, but for a good test, we want the amount of error to be small in comparison with the amount of information—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 all students have a 'true score' on a particular test. This notion of a 'true score' is often misunderstood. Assuming that a student has a 'true score' on a particular test 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 called 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). So, the basis of classical test theory is the assumption that we can write:

$$X = T + E$$

This equation sums up the fact that on a particular occasion, 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). A reliable test is one in which the values of E are small when compared to T, because when E is small, the values of X that we actually observe, or example, for a class of students, are quite close to the values of T, which is, after all, what we wanted to know.

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. In stead, we use a measure of how spread out the values are, called the standard deviation. The key formula is

standard deviation of errors = standard deviation of observed scores

where r is the reliability coefficient, sometimes abbreviated to just the reliability, of the test.

A coefficient of 1 means that the standard deviation of the errors is zero, which means that there is no error, so the test is perfectly reliable. A coefficient of 0, on the other hand, means that the standard deviation of the errors is the same as that of the observed scores—in other words, the scores obtained by the individuals are all error, so there is no information about any individuals' capability in the test at all. If a test has a reliability of zero, this means, in effect, that the result of the test is completely random.

There are several standard methods for estimating the reliability of a given test. One of the first was published by Kuder and Richardson in 1937—the formula for the reliability of a test was the twentieth formula in the paper, and ever since then, this method has been known as KR20.

The reliability of tests produced in schools is typically around 0.7 to 0.8, while commercially produced educational tests have reliabilities from 0.8 to 0.9. Some narrowly focused psychological tests have reliabilities over 0.9. To see what this means in practice, it is useful to look at some specific kinds of tests.

The reliability of standardised tests

Reporting the raw marks achieved by students on tests is not very informative, since we need to know how hard the items in the test were before we can make sense of the scores. Because calibrating the difficulty of items 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 if someone scores 115 on, say, a reading test, we can say that this level of performance would be achieved by 16% (ie half of 32% which is 100% - 68%) of the population, or, equivalently, 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 reading 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, in the accompanying manual), 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

which works out to be just under 6.

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

- For 68% of the students who take this test, their actual scores will be within 6 (ie one standard deviation) of their true scores
- For 96% of the students who take this test, their actual scores will be within 12 (ie two standard deviations) of their true scores
- For 4% of the students who take this test, their actual scores will be at least 12 away from their true score.

What this means is that while for most students in a class, their actual score will be quite close to their true score (ie what they 'should' have got), for at least one student, the score is likely to be 'wrong' by 12 score points. The problem, of course, is that 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 standard error of measurement is 7.5, which means that someone who actually scores 115,leading us to believe that they are in the top sixth of the population might really have a true score of 100 making them just average or as high as 130, putting them in the top 2% of the population (often used as the threshold for considering a student 'gifted').

It is important to note that, because the effects of unreliability operate randomly, the averages across groups of students are likely to be 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, and 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 that 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 (if at all) than 74%, there is a temptation 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.

No data have ever 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 . However, by creating simulations of a 1000 students at a time, it is possible to 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 and this is 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

It is possible to 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 these kinds of changes is small. There are only two really effective ways of increasing the reliability of a test: make the scope of the test narrower, so you can 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

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 proportion of students correctly classified. 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, but 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 possibilities 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 predicting depends entirely on the correlation between the scores on the test used for prediction (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 expect to find that those who got high scores in the IQ tests at age 11 would also go on to get high scores in GCSE, and those who get low scores in the IQ tests at age 11 would go on to low high scores in GCSE. 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’. If someone is interested in the validity of IQ scores as predictors of GCSE scores, what is generally meant by this is the correlation between the true scores of individuals on the predictor and their 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. In some senses this correction is sensible, because there are three stages involved:

the observed score on the predictor is used as a proxy for the true score on the predictor

the true score on the predictor is used as a proxy for the true score on the criterion

the true score on the criterion is used as a proxy for observed score on the criterion

and only the middle stage is really an issue of validity. However, it must be borne in mind that validity coefficients are usually ideal values, and that if we really use these predictors to try to predict performance, then the predictions will be substantially less good than would be suggested by the validity coefficient because of the effects of test 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 for a particular purpose. For example, suppose we wish to 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. We place the 35 students that score highest on our predictor test in set 1, the next highest scoring 30 in set two, and so on. 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. We will place 12 students who should be in set 2 on the basis of their mathematical ability in set 1, and 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 set 3.

Of the 20 students placed in set 3, only 7 should be there—4 should have been in set 4 and 9 should have been in set 2, while of the 15 students we place in set 4, only 7 should be there—four of those placed in set 4 should be in set 3 and four of them should be in set 2! This is summarised 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—if anything, the assumptions made here are rather conservative in that reliabilities of 0.9 and predictive validities of 0.7 are at the limit of what we can achieve with current testing 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	9	
in	set 3		9	7	4
	set 4		4	4	7

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 not much point in being able to measure something reliably unless one knows what one is measuring. After all, that would be equivalent to 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. After all, 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 by the constructors of the mark scheme cannot be given high marks).

In resolving these conflicts, it is important to bear in mind that 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 (cf high reliability), but one has no idea what is going on elsewhere, and the people in darkness, knowing that they are not illuminated, can get up to all kinds of things, knowing that they won't be seen (cf 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 to 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 validityare 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

Kuder, G. F. & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika,* **2**(3), 151-160.

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.