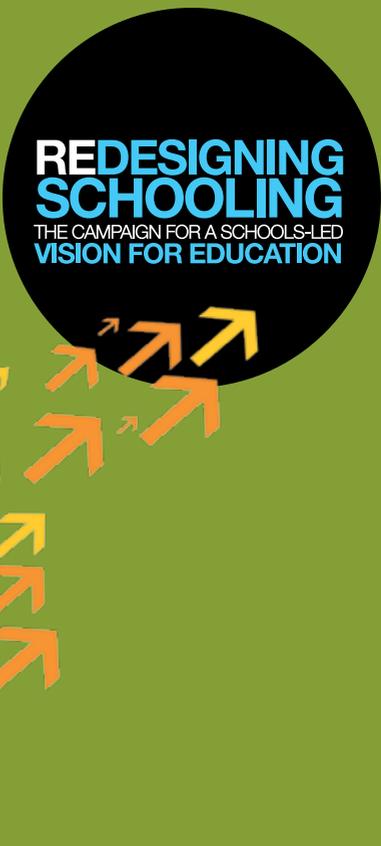Redesigning Schooling - 8

Principled assessment design

Dylan Wiliam
July 2014

REDESIGNING
SCHOOLING
THE CAMPAIGN FOR A SCHOOLS-LED
VISION FOR EDUCATION

Author

Dylan Wiliam is emeritus professor of educational assessment at the Institute of Education, University of London where, from 2006 to 2010, he was deputy director. In a varied career, he has taught in urban public schools, directed a large-scale testing programme, served a number of roles in university administration, including dean of a school of education, and pursued a research programme focused on supporting teachers to develop their use of assessment in support of learning.

His recent work has focused on the use of assessment to support learning (sometimes called formative assessment). He was the co-author, with Paul Black, of a major review of the research evidence on formative assessment published in 1998 and has subsequently worked with many groups of teachers around the world, on developing formative assessment practices. He is co-author, with Siobhan Leahy, of SSAT's *Embedding formative assessment: professional development pack*.

Dylan is also the author of the SSAT title *Redesigning Schooling – 3: Principled curriculum design.*

Editors

Peter Chambers and Jane Birbeck

SSAT's purpose

SSAT believes that teachers make students' lives. As the world gets more complex, that vital role becomes ever more demanding. As the hub of the largest, longest-standing independent network of education professionals in England, SSAT exists to help teachers perform their job even better, more confidently and more professionally than before.

This publication

Audience: Education professionals at all stages and settings, and parents

Aims: This challenging pamphlet comes at an important moment: with the removal of national curriculum levels from September, Dylan Wiliam argues forcefully that schools must develop their own way of assessing children's progress, and that this moment represents an 'extraordinary opportunity' to design an assessment system that is 'the servant, not the master, of the learning'. He guides us through a nuanced way of thinking about assessment that should underpin the design of any model: knowing the limitations of types of assessment and the impact these have on the inferences you make; defining what you will assess and the purpose for which you will use the data; how you will collect, share and/or record it; and, critically, how you can design assessment systems that first and foremost support teaching and learning. 'Good summative assessment requires teachers to share a construct of quality, while good formative assessment requires helping students share the same construct of quality.'

Contents

Introduction

The premise of this pamphlet is that any assessment system should be designed to support the curriculum in place in a school, rather than having the curriculum designed to fit the assessment system. Or to put it another way, assessment should be the servant, not the master, of the learning.

For two decades, assessment in schools has been driven by the eight- (originally ten-) level structure that was proposed by the Task Group on Assessment and Testing (TGAT) in 1987. The model proposed by the task group represented a radical departure from existing assessment and reporting practice, and may well have been different from what the government had expected. It seems that what Kenneth Baker, the secretary of state for education and science at the time, had envisaged was a series of 'benchmarks' at the ages of 7, 11, 14 and 16. What the task group proposed was that all students should be on the same ladder, with a common achievement scale for all students of compulsory school age.

While the model had many strengths, the increasing pressure on schools to improve student achievement on national tests and

examinations, combined with the behaviour of school inspectors, resulted in a situation in which pursuit of levels (or sub-levels!) of achievement displaced the learning that the levels were meant to represent. It is therefore not surprising that the National Curriculum Expert Panel (of which I was a member) proposed that the use of the TGAT assessment model should be discontinued (James, Oates, Pollard, & Wiliam, 2011). This recommendation was accepted by the Secretary of State for Education, and in March 2014, the Department for Education announced that while statutory assessment would continue at the end of each key stage, levels of achievement for assessing achievement during key stages would not be replaced and it would be up to each school to decide how to monitor students' progress towards the expectations for the end of each key stage (DfE, 2014).

Understandably, this decision was far from universally popular. Many have argued that parents were, at last, beginning to understand the system – indeed, many parents had themselves experienced the system at school when they were children. Others have argued that for all its faults, the system allowed achievement in all school subjects to be reported on a common scale. However, against this, it does seem that the idea of age-independent levels of achievement was better suited to some subjects than others, and schools are now free to design assessment systems that meet their needs and the needs of their students.

Formal assessments will continue at the end of primary and secondary schooling, in the form of key stage 2 tests and GCSE examinations, but it will be up to the schools to decide how to get their students to the highest achievement possible in these formal assessments. Schools will therefore be able to plan for the long term. Primary schools will receive children at the age of four or five, and it will be for them to decide what kind of curriculum will suit their local needs best. Secondary schools

will receive students at the age of 11, and be able to plan five-year programmes of study that will allow their students to get the best possible grades at GCSE, and to prepare them for the complex ever-changing world of the 21st century.

Many schools are currently choosing, for the time being at least, to continue with the current levels of achievement, and this may be entirely appropriate in terms of managing workload. But it is also important to note that there will be no straightforward way to carry the existing levels of achievement forward into the new national curriculum, since the new national curriculum will not provide descriptions of the levels. Even when schools choose to use something similar to the current levels of achievement, it may be appropriate to consider whether such levels of achievement are equally appropriate in all subjects. It may also be useful to consider whether dividing levels up into sub-levels makes sense for all subjects. Even where subdivisions do make sense, it may be that two, or four subdivisions may be more appropriate than three, as is currently the case.

The freedom presented by the recent changes is obviously a little daunting, since schools are losing supports on which they have relied for many years – indeed most of our teachers have known nothing else. But the opportunity to design an assessment system that works for the school, rather than the other way round, represents an extraordinary opportunity for schools.

The aim of this pamphlet, then, is to provide schools with some 'tools for thinking' as they begin the task of designing their own assessment system. A school's assessment system has many jobs to do. It must provide useful information for students, parents, teachers and leaders about the progress being made by students. At all times, the system should provide information about important checkpoints or milestones

in learning. But, as students approach high-stakes assessments such as key stage 2 tests and GCSE examinations, it should also provide information about the likely outcomes of those assessments. Some of the information generated by the system will be carefully recorded while other assessment information will be of a more ephemeral nature, informing decisions about next steps in teaching, for example, where formal recording could be unduly burdensome, or even counterproductive. The information also needs to be generated in a timely manner. For some kinds of assessment information, the 'sell-by' date is very close to the collection of the evidence, while other kinds of assessment information has a longer 'shelf-life'. Inevitably, these different demands on the assessment will create conflicts and trade-offs, many of which cannot be resolved easily, if at all. The aim of this pamphlet is not, therefore to tell schools how to design their assessment system, but rather to help them understand better what trade-offs they are making in designing their assessment system, and, in particular, what the costs and benefits of various decisions will be.

In presenting these issues, I have tried to steer clear of jargon and technical issues as much as possible, but the field of assessment is one where it is impossible to understand what is going on without some engagement with the technical details. I have kept this to what I think is the minimum necessary to engage properly with the issues. Inevitably, however, some parts of the pamphlet may appear to be rather like a textbook on assessment.

> In some cases, formal recording could be unduly burdensome, or even counterproductive

It is also important to realise that there is a great deal of science here. By that I mean that there are some issues in which personal opinions

play very little part, and this is rather rare in education. As many people have pointed out (see, for example, Lagemann, 2000) education lacks strong disciplinary foundations, and there is relatively little that could be regarded as 'reliable knowledge' (Ziman, 1978). In the field of assessment, however, there are some very well-understood technical aspects where failure to appreciate the technical details means that one can do things that are just wrong. This is particularly true for the issue of reliability, where schools routinely say things to students or parents that are not just unprofessional (in that they contradict professional standards or codes), but actually incorrect. For example a teacher may state that a student who gets the same level of achievement at the end of the term that he did at the beginning has made no progress. As we will see there is no certainty that this is true.

Chapter 1 begins the discussion by outlining the main reasons we assess student performance in schools, while chapter 2 introduces some features of the design of school assessment systems, specifically what should be assessed. Chapter 3 explores in some depth the issues of the quality of assessment, focusing on the central concept of validity. Reliability, which is a key aspect of validity, is discussed in more detail in chapter 4. Chapter 5 deals with assessment design, and focuses explicitly on the idea that assessments should be designed backwards from the inferences they are designed to support. The interpretation of evidence is discussed in chapter 6 and in chapter 7, I discuss how the results of assessments can be recorded and reported to various stakeholders. Chapter 8 draws the various threads of the other chapters together.

Assessment is a vast field, and inevitably, I have been highly selective in my coverage of aspects of assessment, both in what aspects to include, and what to say about the aspects that have been included. No doubt others would make a different selection, and indeed, had I

written this pamphlet a few years ago, I may well have made a different selection – this pamphlet is very much a product of its time.

In making these selections, I have steered clear of trying to propose a model for a school assessment system – such a model might be suitable for some schools but not for others. Instead, I have tried to focus on some key principles for assessment that will help school leaders and others to think about assessment in a more disciplined way. These principles will often, if not always, be in tension, so there definitely won't be any assessment system that satisfies all the principles. But by making these principles explicit, my hope is that the compromises that are being made are clear and explicit, and above all, planned, rather than just being the unintended consequences of decisions taken by different people, in different parts of the school, at different times.

> I have tried to focus on some key principles for assessment that will help school leaders and others to think about assessment in a more disciplined way

Chapter 1

Why assess?

The word 'assessment' comes originally from the Latin *assidere*, which means, literally, 'to sit beside' someone, presumably in order to form a judgement about their capabilities. As so often, however, the meaning of the word has evolved over the years, and it is now used in very different ways in different contexts.

Even within education, we assess for an extraordinarily wide range of reasons. A speech pathologist might assess a child to find out why the child's speaking is not developing as rapidly as expected. An educational psychologist might give a child some tests of word-processing capability and some more general tests of ability to determine whether the child has a specific learning disability such as dyslexia. And of course, we assess students in school to find out what they have learned.

Because assessments are so prevalent – because they often just seem like part of school life – we often forget that they are conducted for particular reasons. Specifically, assessments are conducted in order to draw conclusions. As Cronbach (1971) pointed out many years ago, an assessment is really just a procedure for making inferences.

This is why a focus on 'data-driven decision-making' that is in vogue in many schools today can be unhelpful; it directs attention to the data that are being collected rather than the decisions that the data will help inform. When we are focusing on the data, a number of potential problems arise.

First, we tend to regard any piece of data as potentially useful – but the result is often large amounts of data for which we have little use.

Second, even when the data are useful, if they have not been collected in a way that supports their immediate use, then it can often be time consuming to get the data in a form that can be used. The result is the information arrives too late to inform the decisions it is intended to support.

Third, when there are data available, even if they are not quite the right data for the decision at hand, the fact that the data are available can result in the data being used just for convenience. Of course this might be sensible, provided that not too much weight is placed on the data. But over time, as the data is routinely used for these decisions, the cautions are forgotten. We start out by wanting to make the important things measurable, and end up making the measurable important.

For these reasons, I believe that it is more productive to focus on decision-driven data-collection, rather than on data-driven decision-making. By focusing on the decisions that need to be made, rather than the data, we are far more likely to collect the right data, the right amount of data, in the right way, for the need at hand.

Sometimes these decisions will be about whether students have learned enough to go on to a future phase of learning, as might be the case with GCSEs or A-levels. But, beginning in the 1990s, there has been increased interest in the UK in the idea that assessment could support education as well as just measuring its results. The use of assessment to improve learning is sometimes called formative assessment or assessment for learning, in contrast to the traditional function of assessment which is sometimes called summative assessment or assessment *of* learning.

As Bennett (2011) has pointed out, assessment for learning is not quite the same thing as formative assessment (nor is summative assessment quite the same thing as assessment of learning), but it is important to recognise that assessment is at the heart of effective learning. As Peter Silcock observed,

> By focusing on the decisions that need to be made, rather than the data, we are far more likely to collect the right data, the right amount of data, in the right way, for the need at hand

> 'We shouldn't want [formative assessment] because research shows how it improves learning (we don't need to be told that – it has to be true). We should want it because schools are places where learners should be learning more often than they are being selected, screened or tested in order to check up on their teachers. The latter are important; the former are why schools exist' (Silcock, 1998).

Another way of saying this is that all assessment should be assessment for learning, as is made clear in the Assessment Reform Group's ten research-based principles for effective assessment:

1. Assessment for learning should be part of effective planning of teaching and learning

2. Assessment for learning should focus on how students learn

3. Assessment for learning should be recognised as central to classroom practice

4. Assessment for learning should be regarded as a key professional skill for teachers

5. Assessment for learning should be sensitive and constructive, because any assessment has an emotional impact

6. Assessment should take into account the importance of learning motivation

7. Assessment for learning should promote commitment to learning goals and a shared understanding of the criteria by which they are assessed

8. Learners should receive constructive guidance about how to improve

9. Assessment for learning develops learners' capacity for self-assessment so that they can become reflective and self-managing

10. Assessment for learning should recognise the full range of achievements of all learners. (Broadfoot et al., 2002).

The challenge, therefore, is to develop assessment systems that can satisfy these laudable aims, while at the same time providing the kinds of information about students' progress that traditional assessments have always supplied. We start by considering what should be assessed.