What is wrong with our educational assessments and what can be done about it.

Dylan Wiliam

*to appear in Education Review **15**(1), Autumn 2001*

Introduction

Students in England and Wales are subjected to more government mandated tests than in any other country, and yet our performance, in comparison with other developed countries, appears to be modest. Furthermore, it appears from the repeat administration of the Third International Mathematics and Science Study (TIMSS-R) that although scores on national curriculum tests are improving at both key stage 2 and key stage 3, overall standards of achievement are getting no better. The proportion of the age-cohort gaining A-level passes is now greater than at any time in our past, and yet employers repeatedly claim that students are less prepared for the world of work than students from previous years. The obvious answer to this apparent paradox is that the tests and examinations have been made easier, but a whole raft of technical studies have shown that standards are being broadly maintained. In this article, I will offer an alternative explanation, which that government initiatives have pressurised schools into improving test and examination scores at any cost, which leads to a narrowing of the curriculum, and which robs tests and assessments of their power to say anything useful about students' achievements. In place of this system, I offer an alternative, based on moderated teacher assessment, with external tasks for public accountability, which ensures that all students receive a broad and balanced curriculum

The curse of quantification and MacNamara's fallacy

For over one hundred years, policy-makers have been searching for objective indices of the quality of education. In the second decade of the twentieth century in the USA, the 'School Survey' movement sought to gather 'objective evidence' about factors influencing the educational progress of school students, but within about twenty years, educational policy-makers looked to psychology to provide a way of measuring the outputs more 'scientifically'. This desire for quantification soon dominated most aspects public-service provision. Perhaps the best-known example of politicians' desire for simple answers to complex questions is John F Kennedy's furious reaction to the ambiguous evaluation of the impact of additional money provided for the education of socioeconomically disadvantaged students: "Do you mean that you spent a billion dollars and you don't know whether they can read or not?".

The trouble with such 'objective' approaches is that while many things can be measured, there are also many important things that cannot, and the danger is that things that can be measured easily come to be regarded as more important that those that cannot. This process is well summed up by Charles Handy's rendering of the what has come to be known as the Macnamara Fallacy, named after the US Secretary of Defense, who argued that the ratio of Viet Cong/North Vietnamese Army losses to US/Army of the Republic of Vietnam losses was an important measure of military effectiveness: "Things you can count, you ought to count. Loss of life is one."

> The Macnamara Fallacy: The first step is to measure whatever can be easily measured. This is OK as far as it goes. The second step is to disregard that which can't easily be measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that what can't be easily measured really doesn't exist. This is suicide. (Handy, 1994 p219)

We start out with the aim of making the important measurable, and end up making only the measurable important.

Lake Wobegon

Education is no exception to these general principles. In 1987 it was discovered that all fifty states in the USA had posted state results on national standardised tests that were above average. This was quickly dubbed the 'Lake Wobegon effect' after Garrison Keillor's mythical town where all the women are strong, all the men are good looking, and all the children are above average. Each state had adopted one of a handful of nationally standardised tests as its 'benchmark', and because there was pressure on teachers to improve their students' scores on these tests, teachers made sure that they taught the material that was going to be tested in the tests. Although this made a mockery of the idea that schools were free to determine their own curricula, the distortions produced were not that great. However, there was a great

range in the results of schools in different areas. While the USA's results in mathematics, for example, were average (and comparable to the UK's) this masked huge variations. In the more affluent areas, achievements were comparable to those in Japan and Singapore, while in the poorer areas they were comparable to Nigeria—one of the lowest-scoring countries in the TIMSS sample. Teachers in the poorer-performing (and poorer) schools were told that they had to improve their results, which, of course, they did, by ensuring not just that the material to be tested was taught, but that this material was the *only* material that was taught. These students were getting better at the material being tested, but nothing else (as was proved when they were tested on other, slightly different standardised tests).

This is the fundamental problem. By making the pressure on teachers and students to achieve good results on particular tests greater and greater, we can secure improvements in scores on those tests, but these improvements are secured at the expense of everything else. The tests, originally meant simply as a sample of the curriculum, come to be the whole curriculum. The reason that this is important is that we are hardly ever interested in the specific things a student has to do to pass an examination or test—after all, a test tests only what a test tests. We are generally interested in examination and test performance because these results can stand as proxies for wider achievement and potential, and in the past. However, by increasing pressure to do well on the test to a ridiculous degree, we have reached a point where we cannot generalise beyond the immediate test scores. When test scores at key stage 2 improve, we cannot conclude that education in key stage 2 has improved. We cannot even conclude that performance in English, mathematics and science has improved. All we can conclude is that the narrow range of skills tested in the key stage 2 tests has improved. This provides an example of what has become known as Goodhart's law.

Goodhart's law

Goodhart's law was named after Charles Goodhart, a former chief economist at the Bank of England, and it states, quite simply, that performance indicators lose their usefulness when used as objects of policy.

The example Goodhart used to illustrate this was that of the relationship between inflation and money supply. Economists had noticed that increases in the rate of inflation seemed to coincide with increases in money supply, although neither had any discernible relationship with the growth of the economy. Since no-one knew how to control inflation, controlling money supply seemed to offer a useful policy tool for controlling inflation, without any adverse effect on growth. The result monetarist policies produced the biggest slump in the economy since the 1930s. As Peter Kellner comments, "The very act of making money supply the main policy target changed the relationship between money supply and the rest of the economy" (Kellner, 1997).

The same problems have beset attempts to provide performance indicators in the Health Service, in the privatised railway companies and a host of other public services. In the NHS, hospitals were told to decrease the number of people waiting two years for an operation, so anyone with a minor condition who had been waiting 23 months went to the head of the queue, and took priority over those with more serious complaints. In some areas, while the target number of people waiting two years for surgery went down, the average waiting time actually went up. On the railways, train operators were fined when their trains ran late, so when a train looked as if it might run late, it was cancelled. The penalties were then extended to reliability as well as punctuality, and then, when trains were running late, they went straight through scheduled stops, just to save time.

If you make a particular performance indicator a policy target, and make the stakes high enough, then the people at the sharp end will do everything they can do improve their score on the performance indicator. However, because the areas in which we use performance indicators are so complex, there is always a way of improving the performance indicator without having any impact on the overall quality of whatever the performance indicator is meant to be measuring (sometimes the quality actually gets worse, even though the performance indicator is rising).

So, when schools were measured by the proportion of students achieving 5 good grades at GCSE, this improved, although in some cases, the average grades achieved by students went down. In response to this the average grades are now also reported, but again schools are able to manipulate this index too, by channeling students towards easier subjects, or by entering students for 'vocational GCSEs' which are deemed to be equivalent to four GCSEs. The reported scores rise, but the actual level of performance may be unchanged, or even declining.

This is the essence of Goodhart's Law—in all these cases, a variety of indicators is selected for their ability to represent the quality of the service, but when used as the sole index of quality, the manipulability of these indicators destroys the relationship between the indicator and the indicated.

There is no end to this process, because the people on the ground will always know more about where the loopholes are than those devising the performance indicators. Put bluntly, the clearer you are about what you want, the more likely you are to get it, but the less likely it is to mean anything.

The fact that our systems of timed written tests and examinations narrow the curriculum is hardly news. However, there is another effect that is less well appreciated, and that is the unreliability of the examinations. There are no up-to-date figures on the reliability of GCSE, but the available data suggest that a student receives the grade that their achievement would merit only around 65% of the time. In other words, around one-third of GCSE grades awarded are wrong. In some cases they are too high, and in others they are too low, and although for many students, the gains and the losses balance out, for students close to the key threshold of 5 A*-C grades, many will miss out not because they weren't good enough, but because they were unlucky.

Tests and examinations can be made more reliable, but the only way of doing this is by making them longer. However, because our tests and examinations are generally such sterile experiences for students, people are complaining that the burden of tests and examinations—up to ten hours for the three subjects tested in national curriculum tests, and up to around fifty hours for GCSE—is already too great. Fortunately, there is a solution.

What can be done?

Our system of tests and examinations distorts our school curricula and produces results that are of limited reliability, and of doubtful validity. In proposing alternatives, the question is not where to find them, but how radical we are prepared to be. Why for example, do students get tested as individuals, when the world of work requires people who can work well in a team? Why do we test memory, when in the real world, engineers and scientists never rely on memory—if they're stuck, they look things up. Why do we use timed tests when it's usually far more important to get things done right than to get things done quickly? There are of course, those who claim that timed written tests give good indications of the ability to work under pressure, in which case, they should produce evidence of this—I haven't seen any. But I have seen plenty of evidence of the damage that timed written tests do, and how poor they are at measuring the important outcomes of learning.

As a modest start, however, accepting the need for formalised assessments of students' achievement at the ages of 7, 11, 14, 16 and 18, I propose that all national curriculum tests (and, if the politicians have the stomach for it GCSEs and A-levels , which is what happens in Sweden, for example) are replaced with moderated teacher assessment. By extending the assessment over the whole key stage, we would produce unprecedented levels of reliability and validity, and the rigorous procedures of moderation would not only ensure against grade drift, but would also provide a valuable focus for inservice training for teachers. This would also be likely to tackle boys' underachievement, because the current "all or nothing" test at the end of a key stage encourages boys that they can make up lost ground at the last minute.

The crucial point, however, in order to prevent teaching to the test, is to disentangle the evaluation of the school from the scores that a student gets. Instead of publishing the results of the moderated teacher assessments, schools would be held accountable by the results of special tasks taken by the students at the end of the key stage. Crucially, there would be a large number of these tasks, and not all students would take the same task. These tasks would cover the entire syllabus, and would be allocated randomly so that there would be no way of teaching to the test. Or more precisely, the only way to teach to the test would be to teach the whole curriculum to every student. Schools that taught only half the curriculum, or concentrated their resources on only the most able students, would be shown up as providing a limited education. Furthermore, the results of these tests could provide an additional check on the robustness of the moderation procedures, and would provide accurate information to policy-makers about the real state of education in our schools.

Summary

Out current educational assessments are not just ineffective—they are preventing us from providing high-quality education for school students, and preventing schools from producing young people with the flexible skills that will be needed in the 21st century.

This is because our assessments started from the idea that the primary purpose of educational assessment is selecting and certifying the achievement of individuals (ie summative assessment)—and have tried to make assessments originally designed for this purpose also provide information with which educational institutions can be made accountable (evaluative assessment). Educational assessment has thus become

divorced from learning, and the huge contribution that assessment can make to learning (ie formative assessment) has been largely lost. Furthermore, as a result of this separation, formal assessment has focused just on the outcomes of learning, and because of the limited amount of time that can be justified for assessments that do not contribute to learning, has assessed only a narrow part of those outcomes. The predictability of these assessments allows teachers and learners to focus on only what is assessed, and the high stakes attached to the results create an incentive to do so. This creates a vicious spiral in which only those aspects of learning that are easily measured are regarded as important, and even these narrow outcomes are not achieved as easily as they could be, or by as many learners, were assessment regarded as an integral part of teaching.

In place of this vicious spiral, I propose that developing a system of summative assessment based on moderated teacher assessment. A separate system, relying on 'light sampling' of the performance of schools would provide stable and robust information for the purposes of accountability and policy-formation.

References

Handy, C. (1994). *The empty raincoat*. London, UK: Hutchinson.

Kellner, P. (1997, September 19). Hit-and-miss affair. *Times Education Supplement*, p23.