National curriculum assessment: how to make it better

To appear in Research Papers in Education (2003)

Dylan Wiliam

(Dylan Wiliam is Assistant Principal and Professor of
Educational Assessment at King's College London)

Address for correspondence:

Department of Education and Professional Studies
Franklin-Wilkins Building
150 Stamford Street
London SE1 8WA

Abstract

In a series of papers ov er the last ten years, I have outlined various problems affecting the assessment of the national curriculum in England which are the subject of a critique by Paul Newton (this issue). In responding to this critique, I acknowledge that his summary of my position is fair, and agree that, by the standards of analytic rationality, the evidence for some of the problems I identify is not compelling. However, in response I argue that by standards of reasonableness (eg on the balance of probabilities) the evidence is sufficently serious to warrant a re-examination of national curriculum assessment, and the alternatives. In particular, I argue that the current system provides assessments that are not sufficiently reliable for the inferences that are made on the basis of the results and has also caused a narrowing of the curriculum. I propose that the first of these weaknesses can be addressed through the increased use of teacher assessment, and the second by increasing the range of the curriculum tested through testing a greater proportion of the curriculum. In order to effect these changes without increasing the burdern on students and teachers, I propose that these two changes are combined in the form of a light sampling scheme which would increase both the reliability and minimise the curricular backwash, although the price paid for this would be the lack of a direct, transparent and objective link between the results achieved by individual students on tests and the reported levels of a school's performance.

Key words

education, national curriculum assessment, reliability, validity, manageability

Introduction

Between 1989 and 1991 I was closely involved in the development of national curriculum assessments for 14 year-olds in English, mathematics, science and technology, and since then I have written on the problems that I believe afflict our national curriculum assessment system (and indeed our examinations at GCSE and A-level). However, I am aware that I have not drawn the various threads of these arguments together, and so I am particularly grateful that Paul Newton has taken the time to trawl through many of these papers (both published and unpublished) and constructed a critique of the ideas I have advanced (Newton, this issue).

In responding to his critique, the first thing to say is that his characterisation of my work in eminently fair and accurate. He has in some places highlighted areas where my thesis was ambiguous or unclear, and therefore in this response I will try to clarify what I was trying to say. He is also right to point out that some of my ideas have been laid out with more rhetorical force that supporting empirical evidence. In most cases, this is because the evidence does not yet exist, lending support to Newton's argument that more evidence is needed. In other cases I have tried to support the thesis, either by additional argument or by citing empirical studies, which while not conducted specifically in the context of the national curriculum of England, nevertheless may be regarded as suggestive.

As far as possible, I have tried to adopt the same sequencing of topics as used by Newton in his paper, although there are places where I have deviated from this in order to avoid repetition.

The validity of national curriculum assessments

As Newton states, I have argued that national curriculum assessments assess only a part of the domain which they are purported to represent. This is partly by design, and partly by accident. By design, the national curriculum tests at key stages 2 and 3 do not assess the first attainment target in mathematics and science nor do they assess Speaking and Listening in English. By accident, or at least without, I think, being planned, the items that do test particular aspects of the national curriculum do so in a distinctive way. For example, in the national curriculum for mathematics, there are requirements for students to collect and interpret discrete and continuous data, which are impossible to assess adequately in a two-hour written test. It is also clear that teachers are able to predict which aspects of a subject do come up in the tests, and which do not (Wiliam, 1993).

Whether the fact that teachers can predict which aspects of a subject are not going to be tested subsequently results in these aspects not being taught is, as Newton notes, an empirical question, but in the absence of appropriate research evidence, I would suggest that the, admittedly less rigorous, evidence from the teaching unions and from school inspection reports presents, at the very least, a case to answer. Indeed, it could be argued that, given the way that narrow targets have distorted performance in the National Health Service and on the railway network, it would be extraordinary were school teaching not so affected.

The third link is the argument is, of course, that if these aspects of a subject are not taught, then the related competences are not developed, which is again an empirical question, and ideally would need to be undertaken for each national curriculum subject.

After all, if it were found that competence in algebra was independent of competence in geometry, this does not mean that we can conclude that competence in reading is independent of competence in writing.

The research that is needed is actually quite straightforward to undertake—my conjecture is that the levels of performance in untested aspects of the national curriculum in each subject should decline while levels in the tested aspects should stay the same, or rise, as has been found elsewhere (Linn, 1994).

The reliability of national curriculum assessments

The main thrust of my arguments with regard to the reliability of national curriculum assessments (and GCSE and A-level examinations) is that data on the reliability of state-mandated assessments should be made available routinely, and that such data should be presented in a form that reflects how the results of the assessments are actually used. In this context, it is worth noting that in the 1970s the examination boards were happy to admit that A-level grades were accurate to at most one grade either way.

In respect of the latter point, I have argued that traditional definitions of reliability, as a form of 'signal-to-noise' ratio designed for continuous variables, create an unwarranted sense of security when used to describe assessments that are reported on discrete scales that are used to support dichotomous decisions. To illustrate this, table 1 (taken from Wiliam, 2000) shows how the reliability of an assessment system looks very different when presented as a classical reliability coefficient and in the form of the number of students getting their 'correct' grades (in the sense of the grade corresponding to their true score) when outcomes are reported on an 8-grade scale .

Reliability	.60	.70	.80	.90	.95	.99
Grading accuracy	40%	45%	52%	65%	75%	90%

Table 1: impact of reliability of marking on accuracy of grading for an 8-grade scale

Newton regards 'misclassification' as a "highly problematic concept", presumably because he regards 'classification' as equally problematic. However, as long as we accept the notion that, for a given assessment, a particular student will have a 'true score' (defined as the long-run average of the scores on repeated takings of the same or parallel tests without learning in between), then a student will have a true level or grade. For students whose true score is close to a level boundary, even if the test is highly reliable (ie yields fairly consistent scores for an individual) then they will sometimes get a level other than their true level. Of course, as Newton suggests, the fact that someone gets 26 marks as opposed to 27 marks doesn't mean much in itself, but, if this means that they get a level 3 rather than a level 4 at the end of key stage 2, it *is* serious—the student may be punished by parents, expectations of the student may be revised downwards, and the student is regarded as in need of remediation in their secondary school, given 'booster' classes and required to repeat the end-of-year 6 test at the end of year 7. Perhaps, with a better understanding of errors of measurement, things would be better, but as long as marks on tests are used to make dichotomous decisions, then I maintain that our measure of reliability should be the accuracy of the decisions.

In a final comment on this issue, Newton suggests that the use of tasks or tests might well result in lower reliability (for the scores for individuals) than with the existing tests —this is absolutely right, but it doesn't matter because these scores are reported and used only at the group level, so that the reliability is close to 100%. Newton points out that the same would be true if the existing tests were reported only at whole-class or

whole-school level, which is also right, and it might well be much better, as has been the case in the US for many years (although is changing rapidly now) for the results to be reported only at school level. However, even if the current tests were reported only at group level, and used to define an envelope of levels that the school could award, this would still create an incentive to narrow the curriculum by teaching only what appears in the test.

Formative and summative functions of assessment

I agree with Newton that one of the strengths of the current system of assessment in place in England and Wales is that the teacher is the student's ally against the external agencies charged with assessment. This makes for a purity of role for the teacher which is attractive. The downside of this, however, is that the failure to use the detailed knowledge that teachers have about their students impoverishes the quality of the summative assessment (and in particular makes it less reliable and diminishes validity). In other words, while teachers may not demand to be involved in summative assessment, good summative assessment demands the involvement of teachers. This is why I believe that we need to find ways of ameliorating the tensions between formative and summative functions of assessment.

My problem with traditional tests is not that they necessitate narrow teaching and rote learning—indeed, our own work with teachers has shown that teachers developing their formative assessment practices produce improvements in learning even when this learning is measured with traditional timed tests and examinations (Black, Harrison, Lee, Marshall, & Wiliam, 2002). Rather, the problem is that such tests do not (or at least do not appear to) *require* deep learning. In high-stakes settings, therefore, teachers may believe that rote-learning provides a short-cut to improved scores. Whether this is true or not is almost irrelevant—there is evidence to suggest that many teachers believe that teaching well is incompatible with improving test scores. Furthermore because the existing tests systematically under-represent the constructs they are purported to assess they create the possibility of increasing scores by increasing a student's competence on only part of the domain.

Incidentally, I have never argued (or believed) that the format of a test item determines the kind of capability that can be assessed, although measuring 'higher-order' skills would appear to be more difficult with multiple-choice items. For reasons that are not entirely clear (and probably not rational) there is a deep mistrust of multiple-choice items in the UK (see Wood, 1991), but in truth, we have never given them a fair trial. We are happy to expend tens of millions of pounds paying markers to mark open-ended items (the total annual marking bill in England across national curriculum tests, GCSE and A-level is around a quarter of a billion pounds), but somehow believe that actually creating the tests should be relatively cheap.

Nevertheless, there are some important differences between what makes a good test item for a formative function and for a summative one. For example, asking students to 'Simplify, if possible, 2a + 5b' (Brown, Hart, & Küchemann, 1984) would be regarded as unfair in a summative test. The expectations of students that one has to 'do work' to get marks in a test might pressure some students (who would otherwise say that this expression cannot be simplified) into attempting to simplify the expression. These kinds of 'trick' questions are generally regarded as inappropriate for high-stakes tests. However, such items provide highly useful information for the teacher and so would be entirely appropriate for a formative purpose. The crucial feature of the system that I propose is that there is no routine aggregation from the teacher's day-to-day records,

kept primarily for formative purposes, to the summative level that would be reported to students and their parents. The teacher would be free (indeed would be expected) to discount evidence related to 'trick' questions like the one given above, when arriving at a level.

The overall profile of levels for a class would be 'moderated' by the external tasks and tests (see below), which would ensure that the levels awarded could not be inflated by the teacher. There are many ways in which this could be done—the most severe would be to use the results of the external tasks and tests to define an 'envelope' of levels that the teacher was allowed to award, so that the distribution of the levels in the summative levels given by the teacher would have to be exactly the same as that for the external tasks and tests. In addition, in order to check that the teacher's weighting of various aspects of the domain was something similar to those intended in the curriculum, requirements for correlation could be imposed, so that, to some extent at least, those getting high marks on the tasks and tests would be awarded high levels. However, this would be a crude measure, and there is no doubt that additional ways would be needed to detect and, where possible, eliminate the forms of bias noted by Newton (eg over-emphasis on certain aspects of the domain, and inclusion of construct irrelevant variance, such as halo effects).

As I note in Wiliam (2000a), care must also be taken to avoid the teacher's role in summative assessment driving underground formative evidence (eg when students do not divulge difficulties to the teacher because they believe it will be 'held against them'). Ultimately, this can only be resolved through trust, but it can be ameliorated through the depersonalisation of the assessment procedure—while the assessment of the student against the criteria is undertaken by the teacher, it is important that the student understands that the criteria themselves are not determined by the teacher, but are external. Although not perfect, the teacher could then still claim to be the student's ally.

Newton also raises the question of whether teachers' assessments would, as I have claimed, be more reliable than those arising from tests. He is right to point out that continuous assessment over the period of the key stage is not a replication of the final assessment, and it would, indeed, be invidious if a student's level were reduced by the teacher because the last recorded evidence of a particular aspect of the domain dated from the previous year. However, if we adopt the conceptual framework provided by generalisability theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972), then it is a logical necessity that the degree of unreliability contributed by student-task interactions will be lower than for a traditional test because there are more tasks, provided, of course, teachers can apply the correct standards, and base their levels on the 'latest and best' evidence. If we can control the other sources of unreliability (task-rater interactions, student-rater interactions, etc—see below) then teachers assessments will be more reliable than tests.

Evaluative assessments

The fundamental feature of the assessment system I propose is that the evaluative function of the assessment is based on light sampling. The logic of this is straightforward. In order to avoid the possibility of 'teaching to the test' we need to assess a greater proportion of the domain of interest. More precisely we want to create a situation in which we are happy for teachers to teach to the test, because the only way to improve the test score is to improve the performance of students on the whole domain.

Newton seems to suggest that this could be achieved by adding 'authentic' elements to the existing assessment, as happens in some science examinations which involved a

practical element. However, the problem with such an approach is the impact of task-student interaction—put simply does that particular practical task suit the student? The work of Rich Shavelson, Bob Linn and others (Linn & Baker, 1996; Shavelson, Ruiz-Primo, & Wiley, 1999) shows that not until the student attempts six or more such tasks does the average score across the tasks provide adequately reliable indications of the student's capability. We could therefore require each student to take at least six tasks but this would be extremely time-consuming. More importantly, it is also unnecessary. The purpose of the external assessment is to ensure that there is no advantage to the teacher or the students of teaching only some of the curriculum, or teaching to only some of the students. However, we can get the same assurance by light sampling, since the teacher will not know which students will be tested on which parts of the curriculum. Now of course the score that a particular student gets will not be a reliable indicator of their achievement because of the task-student interaction described above, but the average score of the class across the particular tasks taken will be an accurate estimate of the average score of the class across all possible tasks (and one whose accuracy we can judge precisely).

This does raise problems of comparability as Newton notes, but the nature of comparability raised by a light sampling approach is not the same as that in which the test is to be used to impute scores to individuals. The classic definition of test equivalence is that two tests are equivalent (ie comparable) if it is a matter of indifference to the candidate which test is taken (Lord & Novick, 1968). For a light sampling scheme, assessments would be comparable to the extent that it was a matter of indifference which particular allocation of tasks or tests to students was actually administered. Of course there will be particular allocations where, by chance, each student in a class is allocated the particular task or test that suits them best, but these will be extremely rare. With such a scheme there would be no requirement for each task or test to be strictly comparable to the others, in the same way that two tests can be equivalent without an item-by-item equivalence. The reliability of the system would, of course, have to investigated, but this could easily be undertaken by the allocation of different sampling schemes to the same classes.

The marking of these tasks would, as Newton notes, be more complex than current practice. It would not make sense to have one marker marking all of a school's tasks and tests because they would need to become familiar with the marking scheme for every task and test. It would be much more sensible to send all the responses for a particular task or test, from many schools, to one marker. While this sounds administratively complex, with the use of bar-coding, this could be accomplished relatively straightforwardly.

The number of tasks and tests that would be required would need further research to determine, but, as Newton states, it would need to be very large to allow the re-use of tasks and tests from year to year. In cases where factors affecting the difficulty of items were well understood, item-shells might be used with computers to generate large sets of items of similar difficulty, and most, if not, all of the tasks and tests could probably be administered by computer within the next few years. Ultimately, even the marking of open-ended items may be possible by computer. However, Newton is right to sound cautions regarding the availability of people to design the tasks and tests, and it would be several years before a large enough bank of tasks and tests could be built up.

Conclusion

No assessment system can do everything, and therefore it is futile to ask "Is this system perfect?" The answer will always be no. What we can, and should, ask, is "Are the trade-offs between reliability, validity and manageability that we have settled upon the right ones?" In particular, we should ask whether the system that we currently have is the only way of satisfying the design requirements.

The current system is transparent in that there is an apparently objective relationship between the scores that students get on tests, the levels they are awarded as a result, and the scores of schools. The position of a school in a performance table is directly determined by aggregating the marks achieved by its pupils in tests. The question is then what are the trade-offs in the current system, and whether there are other ways of achieving the same ends with fewer adverse consequences. While there is no conclusive system-wide research to demonstrate that the adverse consequences of the current system are serious and far-ranging, I would maintain that there is enough evidence to suggest that something is seriously wrong with the current system. In doing this, I am arguing, along with Toulmin (2001) that in the absence of reliable knowledge about a particular issue, we sometimes have to rely on what appears to be reasonable. In the case of national curriculum assessment, while it may not be possible to demonstrate the adverse consequences 'beyond a reasonable doubt', I believe the case *is* established 'on the balance of probabilities', particularly in terms of curricular distortion and the consistency of levels attributed to individual pupils.

In outlining an alternative model of national curriculum assessment my concern has been to attempt to work towards a system of assessment that delivers the same outputs as the current system—measures of the achievement of individual students, together with evaluative information on schools—but with fewer adverse consequences.

The light sampling approach that I have outlined certainly could be expected to reduce the incentives to teachers for narrowing the curriculum, and may increase the reliability of the assessments made of individual pupils, although, as Newton notes, these are empirical questions which could be settled by undertaking further research. The trade off would be a lack of transparency in that the levels awarded to pupils would derive from integrative judgements by their teachers, rather than by the aggregation of marks.

Ultimately, the differences between Newton and myself seem to me to be mostly about the burden of proof. He regards the arguments I have advanced regarding both the deficiencies of the current system, and the strengths of my proposed alternative, as 'not proven'. In the final sections of his paper, he lays out the essential elements of a research agenda, both into the adequacy of the current system, and of the alternatives. This is a very helpful contribution, and the challenge to research the researchable questions needs to be taken up. But at the same time, I think it is fair to ask whether we must we wait until all the evidence is in before things change. There is always the danger of making things worse, captured in the old adage that we cannot countenance change—things are bad enough as they are! The challenge for the educational research community is to provide policy-relevant findings when we cannot be certain about what to do. I welcome Newton's contribution, not least in forcing me to clarify and develop my own thinking, and hope that others, too will join in this debate.

References

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: assessment for learning in the classroom*. London, UK: King's College London Department of Education and Professional Studies.

Brown, M. L., Hart, K. M., & Küchemann, D. (1984). *Chelsea diagnostic mathematics tests: algebra*. Windsor, UK: NFER-Nelson.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioural measurements: theory of generalizability for scores and profile reporting*. New York, NY: Wiley.

Linn, R. L. (1994) *Assessment-based reform: challenges to educational measurement.* Paper presented at Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.

Linn, R. L., & Baker, E. L. (1996). Can performance-based student assessment be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based assessment—challenges and possibilities: 95th yearbook of the National Society for the Study of Education part 1* (Vol. 95(1), pp. 84-103). Chicago, IL: National Society for the Study of Education.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on sources of sampling variability in science performance assessments. *Journal of Educational Measurement, 36*(1), 61-71.

Toulmin, S. (2001). *Return to reason*. Cambridge, MA: Harvard University Press.

Wiliam, D. (1993). *Technical issues in the development and implementation of a system of criterion-referenced age-independent levels of attainment in the National Curriculum of England and Wales.* Unpublished PhD thesis, King's College University of London.

Wiliam, D. (2000, November) *Integrating summative and formative functions of assessment.* Paper presented at First annual conference of the Association for Educational Assessment-Europe held at Prague, Czech Republic. London, UK: King's College London School of Education.

Wiliam, D. (2000b). The meanings and consequence of educational assessments. *Critical quarterly, 42*(1), 105-127.

Wood, R. (1991). *Assessment and testing: a survey of research*. Cambridge: Cambridge University Press.