English Teaching: Practice and Critique December, 2008, Volume 7, Number 3
http://education.waikato.ac.nz/research/files/etpc/2008v7n3art5.pdf pp. 65-73

The end of testing and future possibilities: An examination of the demise of Sats in England and possible alternative assessments

BETHAN MARSHALL
King's College, London

ABSTRACT:
This article looks at the controversial starting of testing, its boycott, the subsequent years of protest and, in October 2008, the apparent end of key stage examining in England. It considers a possible alternative to the tests based on a project carried out at King's College London based on portfolio assessment.

KEYWORDS:
Key stage testing, portfolio assessment, guild knowledge, construct referencing, course-based assessment

A kind of miracle has occurred within the teaching of English in England. Until recently the government was going to impose a vast array of new tests on pupils and teachers alike. Particularly, they were going to replace or add to the national curriculum key stage tests at eleven and fourteen. The key stage two and three exams gave a global figure of what level of attainment a pupil had achieved according to national curriculum levels – typically at eleven they achieved a level 4 and at fourteen a level 5. But a new set of exams, called single-level tests, was piloted. They were so-called because pupils were to sit them every time their teacher felt that they had arrived at a new level of achievement. The opportunity to take these new tests was to be twice a year and the results were to be just as high-stakes as the previous key stage tests. A bright pupil, for example, might take three tests instead of one at key stage three – a level test for levels five, six and seven.

Then the unprecedented happened. In the middle of an economic crisis, amongst the *sturm und drang* of buying banks and the plummet of the pound against international currencies, Ed Balls, the Secretary of State for Children, Schools and Families, announced that the key stage tests as we had known them and loathed them, were to be no more. Not only that, he abolished the single-level tests for key stage three and has questioned them for key stage two as well. So at the moment we in England live in a halcyon phase, where the dreaded tests for fourteen-year-olds are no more, and nothing as yet has replaced them and everything is up for grabs at key stage two. All this may change when the committee reports on its findings on what to do at both key stages. We have until March 2009.

Before we celebrate too much, however, we must still note that the functional skills tests, to be taken at GCSE, remain and the picture of English they paint is at best reductive and at worst positively harmful but they must be written of in another article. This article then, which was to be a complaint against the overwhelming nature of testing in England, has a slightly different slant, part historical, part empirical and part ever hopeful.

Copyright © 2008, ISSN 1175 8708

A BRIEF HISTORY OF NATIONAL CURRICULUM TESTING

We must start, therefore, by looking at what these key stage tests were supposed to be about, for they still may occur at key stage two. When the national curriculum arrived in 1989 it announced that three subjects – Maths, English and Science were to be assessed at three key stages – seven, eleven and fourteen. Pupils were to be given a level for the standard that they had achieved – the rationale being that one could tell how they had progressed throughout their school career. The levels began life as separate statements, Statements of Attainment, as to what a child could do, but in English, particularly at secondary level, teachers were encouraged to look at these statements holistically. These statements only lasted a very short time and by the second national curriculum of 1995, each level had been replaced with broad characteristics of achievement. Although they had a technicist flavour – the levels for presentation, punctuation and spelling having been incorporated into an overall level for writing – it was still possible to interpret them in a way that was broader than simple standards of literacy.

To return to the beginnings of the key stage tests, however, the Conservative government, who was in power at the time, began with the tests for seven-year-olds while still piloting the tests for fourteen-year-olds. This was a pity. While the pilots for the key stage three tests were going quite well, in English at least, the exams for seven-year-olds were a disaster, and the whole testing regime ground to a halt. The pilots for English key stage three, which had viewed the levels holistically, were rudely interrupted, the people who had the contract (the CATs consortium) were sacked and a new contract was issued, this time to the Northern Examinations and Assessment Board (NEAB). Instead of being tasks at fourteen, which pupils were asked to do over a period of time, pencil and paper tests were introduced; an anthology of literature and a Shakespeare play was added to the items to be assessed. A furore broke out, led by the English teaching profession, which eventually led to the tests being nationally boycotted in 1992 and 1993.

After two years, however, things broke down and the first key stage exams were taken in 1994. The English tests now assessed comprehension, writing and, at fourteen, a Shakespeare play. There was no test, at any age, for speaking and listening. Numerous complaints ensued about the new regime, particularly from English teachers, that the tests restricted the curriculum. The London Association for the Teaching of English (LATE) collected many of the objections that teachers felt about doing the so-called Sats exam. The first report was actually collated during the national boycott; the second after pupils had done them for one year. They were called *Voices from the classroom* (1993) and *The real cost of SATs* (1995). As one teacher wrote in the in the second booklet:

> The Sats have a negative influence on the curriculum because they narrow and limit what can be done. They tend to eliminate creativity and imagination in both the teachers and the student. Instead we are told what to do, what play to read, and what scenes will be examined (LATE, 1995, p. 31).

While these are just anecdotal accounts, the teaching unions carried out surveys of teacher opinion. The National Union of Teachers (NUT) completed a questionnaire of secondary English, Maths and Science in the summer of 1995 (Close, Furlong & Swain, 1995). It asked to what extent teachers felt the tests had altered their practice and also to what extent they used past papers. By far the highest number to believe that their practice was being affected were English teachers. Instead of being able to encourage a wide range of responses to a text, they felt that the potential for variety posed a threat which needed to be circumscribed by close attention to the likeliest answers.

Research carried out by the Association of Teachers and Lecturers (ATL, 1996), published the following year, found that nearly three-quarters of English teachers felt they had been "teaching to the test more than was reasonable" (ATL, p. 16) as opposed to just over a quarter of Maths teachers and almost a fifth of Science teachers. In fact, the figures also showed that while English teachers had become increasingly frustrated with the distorting effect of the tests, Science and Maths teachers were accommodating more easily. The figure of dissatisfaction for English teachers stood at just over half, compared with over around two-fifths of Maths teachers and a fifth of Science teachers. Nearly 80% of English teachers felt that "the tests [had] narrowed the curriculum" (ATL, p. 12).

Arguments about the validity of the tests raged throughout their fourteen years on the books. Every summer the papers were furnished with complaints that the Sats had failed to show the ability of pupils, particularly the most able; that they often gave high scores to pupils who did not deserve them, or that the results were a real muddle. Every year schools sent back copies of the tests so that they could be remarked, but the government, of whatever complexion, stood by them.

When Wales, however, achieved a degree of independence from Westminster, it undertook a study of the testing arrangements for pupils in its schools. Professor Richard Daugherty was placed in charge of the group, which reported in 2004. In an interview about the recommendations of *Learning pathways through statutory assessment: Key Stages 2 and 3*, he concluded:

> What the Group has tried to do, by reviewing all the evidence we could find and taking the best available advice, is to learn from the experience of the past ten years. We have concluded that some features of the current arrangements, such as assessment by teachers at the end of each key stage, should be retained but strengthened. Other features, such as the core subject tests taken at the same time as teachers are making their assessments, should be phased out (Docherty, 2004).

Scotland had never done key stage tests and now Wales was to join them. In Wales, the tests were to be replaced by an APU (Assessment and Performance Unit) style assessment for fifteen-year-olds, so that international comparisons might be made, and skills tests were to be introduced for the penultimate year of primary schools. Work in the final year of primary school was to be moderated by teachers in primary and secondary schools in the same education authority. Only pupils in England now did the tests.

In July 2008, the House of Commons Select Committee's on Children, Schools and Families, reported on exams in England. The select committee had, for a number of

months, asked those involved in any way with testing their opinions and research findings on the assessment regime in England. The report was wide-ranging, but in it the committee found room to criticise the tests as they currently stood. On the purpose of national testing, for example, they wrote:

> We consider that the overemphasis on the importance of national tests, which address only a limited part of the National Curriculum and a limited range of children's skills and knowledge, has resulted in teachers narrowing their focus (Great Britain Parliament House of Commons Children, Schools and Families Committee, 2008, p. 92).

And again, "We are concerned about the Government's stance on the merits of the current testing system" (p. 92). On the consequences of high stakes testing, they wrote: "We believe that teaching to the test and this inappropriate focus on test results may leave pupils unprepared for higher education and employment" (p. 94), saying also that they were "persuaded by the evidence that it is entirely possible to improve test scores through mechanisms such as teaching to the test, narrowing the curriculum and concentrating effort and resources on borderline students" (p. 94). It was on the single-level tests, however, that they were most emphatic. "We believe that true personalised learning is incompatible with a high-stakes single level test which focuses on academic learning" (p. 96). Although those involved in education took note of these findings, however, the government seemed not to heed them.

The final nail in the coffin, in England for the testing arrangements at key stage 2 and 3 seemingly came somewhat later that summer. For the first time the contract for marking the Sats had gone to ETS (Education and Testing Service), the American company who examine the SATs in the US. The arrangements were a fiasco from start to finish – marking the papers wrong and failing to assess all the candidates were amongst the disasters – and eventually their £156m, five-year contract was severed in August 2008.

At the time, the British Government said that it was going to award a single-year contract to a new body. But it seems that even the government realized that this was a mistake. Mary Bousted, the president of the ATL, had called for "The government… to take this golden opportunity to completely overhaul its testing regime, and in the interim should suspend SATs for 2009" (quoted in Lipsett, 2008).

On October 14, Ed Balls went one stage further and got rid of the key stage three tests altogether as well as the single-level test in secondary schools and potentially the tests for eleven-year-olds. He argued that the main indicator of how a secondary school was doing was the GCSE; that the key stage three tests had served their purpose but were no longer necessary and that the single level tests did not differentiate enough between pupils to make them worth pursuing. At key stage two, however, his sentiments were somewhat different. He said that the tests were "essential to giving parents, teachers and the public the information they need about the progress of every primary age child and every primary school" (Balls, 2008). They were to continue to evaluate the single-level tests.

In order to do this, Ed Balls commissioned a committee to look into the testing arrangements at both key stages two and three. The committee is to report at the end

of February, 2009. Although nobody in the group is directly involved in assessment, they have two advisors, one of whom is Professor Dylan Wiliam, well known for his work on formative assessment. This is where we now stand and it is into this gap that I would like to inject a brief piece of optimism. I say brief, because all could be rejected come March.

KING'S OXFORDSHIRE SUMMATIVE ASSESSMENT PROJECT

In 2002, King's College London undertook research into assessment at key stage three in Maths and English. The research was funded at first by the then DfES, and, after a year, for two years by the Nuffield Foundation. We hoped to consider course-based assessment as a way of deciding the national curriculum level awarded at key stage three. It is this strategy that I hope might become the new system for both key stage two and three. As such it is worth considering the project in slightly more detail, concentrating on English.

Essential to the project was a desire for both reliability and validity in both subjects. The key stage tests had placed at the centre of their concern the reliability of the tests. They were nationally accountable tests of a child's achievement and as such placed the emphasis on whether or not the tests could be marked and remarked in the same way again. The result was that while much attention had been given to their reliability, little concern had been given to their validity. This, as we have seen, was the chief complaint against the tests from English teachers since little or nothing had changed in the intervening years.

Because the Sats were still in existence at key stage three, we decided to look at whether or not it was possible to develop both a reliable and valid system of assessment one year earlier in the curriculum. As the tests were taken in year 9 by fourteen-year-olds, we concentrated our study on thirteen-year-olds in year 8. We undertook to work with three Oxfordshire schools that were known for their work in formative assessment, working with the head of department and one other member of staff. We met the teachers for regular meetings and interviewed them twice.

Several studies of course-based assessment had taken place both before and after the SATs, the most comprehensive being the EPPI study (Harlen, 2004). While this gave a somewhat confused picture of course-based assessment, it found two important features:

1. If there was any success in teacher's summative assessment it lay amongst English teachers;
2. Course-based assessment took time to establish.

The Queensland system of course-based summative assessment, however, was clearer. It was wholly school-based, aiming for criteria-led assessment with "respect for teacher professionalism in judging student achievement" (Cummings & Maxwell, 2004, p. 93). This echoes the mode of assessment, finally abandoned in 1994 by the then Tory government, for the 100% coursework examinations for English at GCSE, by what is now the AQA.

Each of these systems, the Australian and AQA model, relied heavily on active teacher involvement at all levels, thus building a community of teacher assessors. Both had first schools and then expert teachers moderating the pupils' work. The AQA, which around 80% of schools in England now take, have maintained their system of standardising and moderation to the present day at GSCE, although the course-based assessment is less than it used to be. The Queensland system and the AQA model are based on many years of research (Petch, 1967; Rooke & Hewitt, 1970; Smith, 1978; Maxwell, 2004; Cumming *et al.,* 2004), which showed that constant meeting and discussion of coursework led to a raising of professional standards and an agreement on grades.

Wiliam (1998) refers to this process as one of construct referencing. Neither norm referencing nor criteria referencing in their strictest senses, construct referencing has elements of both and something else besides. It took a practice of impression-marking, common amongst English teachers, one stage further. In essence, when English teachers award a grade to a piece of work, or a folder, they are using a construct of what they think that grade looks like, based on their previous encounters with work of a similar standard.

Sadler's (1989) use of the term *guild knowledge* is also relevant. He has argued that teachers use a sense of what it means to be very good at something and translate such knowledge, "guild knowledge", into everything they mark. The final grade awarded mirrors how closely you approximate what it means to be good at that particular thing. Where formative assessment is involved, the teacher encourages the pupil, through peer marking and the like, to enter that guild.

In essence, we used both Wiliam's model of construct referencing and Sadler's idea of guild knowledge with the KOSAP teachers. In particular, we asked them what it meant to be "good" at English. When interviewed, English teachers felt, particularly at key stage three, that external tests lacked validity in that they did not capture what it meant to be good at the subject. During their interviews and project days, they equated quality with, "insight", "flair" and "confidence" amongst other attributes. They felt that someone who excelled in the subject showed interest, enjoyment and engagement with language that allowed for risk-taking and "subverting conventions". In other words they saw English as a language art, not a study in communication.

When we started the project, each of the schools involved already kept a portfolio of pupils' work at key stage three and very little alteration took place over the two-year period. What changed was that the requirements became standardised and more rigorously applied, and speaking and listening were added. The portfolio, then, included three reading, three writing and three speaking and listening assignments, but they overlapped, so that there was an assignment where reading and writing were assessed together and one where reading and speaking and listening were jointly marked.

During the course of two years, we held two standardisation meetings and one final moderation meeting. At the moderation meeting, schools had already blind-marked their own candidates' work and come to an agreement as to the final grade they were awarding. A sample of work from each school was then sent to the two other schools involved, also to be blind-marked. Each school then arrived at the final meeting

knowing their own grades and having blind-marked a sample of the other two schools. A moderation meeting then took place where all the grades were discussed.

What was striking was that while there was disagreement as to some of the levels which were awarded, agreement was found in the end. It seems that the KOSAP teachers had begun to agree on what a construct of a key stage level looked like. The only difficulty lay with the speaking and listening grades. It was found that it would be easier if GCSE criteria rather than national curriculum levels could be used. It is possible that this is because speaking and listening has long been a part of GCSE criteria but has never been examined at any of the key stages in the SATs tests.

But the GCSEs may have had a greater influence, or impact, than on just speaking and listening. The key stage tests acted like a kind of hiatus on progression in English. As one teacher put it, "You prepare them in a very specific way and you boost them and you give them strategies and you programme them to do things in a certain way. And that's not the way I would naturally teach." It seemed that GCSE was considered a better indicator than the Sats as to what it meant to be good at English in general, so that English teachers involved in the project began to think about what GSCE grade, as well as NC level, they would give candidates. This meant that they started to think about the progress of a pupil holistically and, in effect, backwards from GCSE. While the national curriculum levels had once been transferable to GCSE – the CATS consortium having built on this – it had become hazier. A level 6 used to be about a level D and a 7 was a C. Over time, though, this had changed and a level 7 had become more like a B. Because they were reported differently and the KOSAP teachers were thinking of levels and then grades, the change had not really become apparent.

As the English teachers were now looking at both levels and grades, it was discovered that in effect some pupils were getting the same mark for SATs and GCSE and they had made little or no progress. Doing the project meant that this process was reversed and teachers began to think in year 7 what they would get at GCSE and start teaching towards it rather than key stage 3 first and GCSE as a separate entity altogether. They were better able to do this because the course work fed into this process and this influenced the way in which they taught.

The net result was that changing the assessment at key stage three affected not only key stage three assessment but GCSE as well. It loosened up the whole system and so changed classroom practice in a beneficial way. As one teacher said, "So you're pulling away the scaffolding aren't you in its various forms and off they go, which is what you as a teacher want to do." Another commented on the influence it had had on GCSE and formative assessment:

> [It's] influenced our GCSE thinking as well, because giving them a lot of scope, but also giving enough support to the ones who need the support… And then creating scope so you know, having open-ended success criteria, getting them to design the task and the success criteria themselves.

The KOSAP study, then, was in part a success. It was found that English teachers could agree, when moderating student grades, and could come to some understanding

of grades for speaking and listening if they applied GCSE criteria. They had a construct of what the levels were. Moreover, considering pupils' progress as a whole improved the quality of learning undertaken at high school. Cheating, another complaint of course-based exams was not seen to be a problem. The teachers all claimed they knew their candidates' work.

For this reason we have recommended to the group studying a replacement for Sats that they look at the KOSAP-type model for English at both key stages two and three. Because the key stage three requirement is not as significant as the key stage two exam, GCSEs being the main and acknowledged marker of achievement at secondary school, we have suggested that portfolio of student work be moderated only at school level. Moderation of the kind done in KOSAP, however, could be done if it was felt that the results had to be reliable and valid and nationally accountable.

But assessment at key stage two would have to be done externally, as well as internally, moderated because it would potentially be a higher stakes exam. A sample of pupils' work would have to be sent in for external moderation, rather as was done in the KOSAP model. This could either be done within or across local authorities. It would also be important, as again was done in KOSAP, to have separate days for standardising material. A sample of work could be sent in – some of which would have to be borderline, that is on the borders of a 3/4 or 4/5. Single-level tests, however, would be abolished. This would be, in part, as has already said before, because children would have to take these high stakes tests more frequently but it would also be because it would interfere with the more holistic approach of portfolio work.

It might be said that teachers are not ready for such extreme changes. It is true that the teachers on our KOSAP study took some time to implement the changes they ultimately felt were necessary, but until we alter our testing system, there is no reason to change at all. In a few months time, we will know if this has just been wishful thinking on our part. A kind of false celebration. But just for once it has been good to dream of a system that could work, not in an ideal world, but a real one that has tired of Sats and single-level tests and very much wants something better.

REFERENCES

Association of Teachers and Lecturers. (1996). *Level Best revisited: An evaluation of the statutory assessment in 1996*. London: ATL Publications.

Balls, E. (2008). Major reforms to school accountability including an end to compulsory national tests for fourteen year olds. More support in year 7 to help children make the jump to secondary school. London: Department for Children, Schools and Families. Retrieved December 8, 2008 from http://www.dcsf.gov.uk/pns/DisplayPN.cgi?pn_id=2008_0229

Close, G., Furlong, T., & Simon, S. (1995) *The impact and effect of KS3 tasks and tests on the curriculum, teaching and learning and teachers' assessments: A report from King's College London University commissioned by the NUT*. London: National Union of Teachers.

Cummings, J., & Maxwell, G. (2004). Assessment in Australian schools: Current practice and trends. *Assessment in Education, 11*(1), 89-108.

Daugherty. (2004). Jane Davidson receives Daugherty group's final report. Retrieved December 8, 2008 from http://new.wales.gov.uk/news/archivepress/educationpress/edpress04/706492/?lang=en

Great Britain Parliament House of Commons Children, Schools and Families Committee. (2008) *Testing and assessment: Government and Ofsted responses to the Committee's third report of session 2007-08, fifth special report of session 2007-08. House of Commons papers 1003 2007-08.* London, HMSO.

Harlen, W. (2004). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Maxwell, G. (2004). Progressive assessment for learning and certification: Some lessons from school-based assessment in Queensland. Paper presented at the third conference of the Association of Commonwealth Examination and Assessment Boards redefining the roles of educational assessment. Nadi, Fiji.

LATE. (1993). *KS3: Voices from the classroom.* London: LATE.

LATE. (1995). *The real cost of SATs: A report from the London Association for the Teaching of English*. London: LATE.

Lipsett, A. (Aug 15th, 2008) US firm loses contract after SATs exam fiasco. *The Guardian.* Retrieved December 8, 2008 from http://www.guardian.co.uk/education/2008/aug/15/SATs.schools.

Petch, J. (1967) *English language: An experiment in assessing. Second Interim Report*. Manchester: Joint Matriculation Board.

Rooke, H. & Hewitt, E. (1970) *An experimental scheme of school assessment in Ordinary Level English language: Third report.* Manchester: Joint Matriculation Board.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144.

Smith, G. (1978). *JMB experience of the moderation of internal assessments.* Manchester: Joint Matriculation Board.

Wiliam, D. (1998). Construct referenced assessment of authentic tasks: Alternatives to norms and criteria. Paper presented at the 24th Annual International Association for Educational Assessment (IAEA) Conference, Barbados, West Indies.

Manuscript received: November 29, 2008
Revision received: December 12, 2008
Accepted: December 14, 2008