DOCUMENT RESUME

ED 347 182 TM 018 602

AUTHOR Gipps, Caroline
TITLE National Testing at Seven: What Can It Tell Us?
SPONS AGENCY Economic and Social Research Council, Lancaster (England).
PUB DATE Apr 92
CONTRACT 000-23-2192
NOTE 14p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; Educational Attainment; *Elementary School Students; Evaluation Methods; Foreign Countries; *National Competency Tests; National Programs; Political Influences; Primary Education; *Standardized Tests; Student Evaluation; *Testing Problems; Test Reliability; Test Validity; Young Children
IDENTIFIERS National Curriculum; *Performance Based Evaluation; *United Kingdom

ABSTRACT

The use of performance based evaluation on a national scale with 7-year-olds in the United Kingdom is described, and the impact of national assessment on teaching practice and implications of this type of assessment are considered. The national assessment program in the United Kingdom started in earnest in 1991 when all 7-year-olds were assessed by their teachers and by external tests, the Standard Assessment Tasks (SATs). The SATs are authentic performance assessments in English, mathematics, and science covering the first three levels of attainment in the national curriculum system in the United Kingdom. The impact of this assessment is being studied in a 32-school sample in a range of schools around the United Kingdom. The SATs represent authentic/performance assessments, and by and large they match the active process-based tasks that children accomplish in good elementary education. These tasks give teachers direct feedback and provide pointers toward a wider view of teaching and learning. Some problems of validity, reliability, and testing length exist. Other problems arise from the complexity of the underlying curriculum structure, the inappropriateness of the model for what was a national survey, and a harsh political climate for a new assessment. However, these problems should not deter the search for quality national assessment. Eight references are provided and one figure illustrates the discussion. (SLD)

National Testing at Seven: What can it tell us?

Paper presented at the
American Educational Research Association Annual Meeting,
1992, San Francisco.

Dr Caroline Gipps,
Curriculum Studies Department,
University of London,
Institute of Education,
20 Bedford Way,
London WC1H OAL.

National Testing at Seven: What can it tell us?

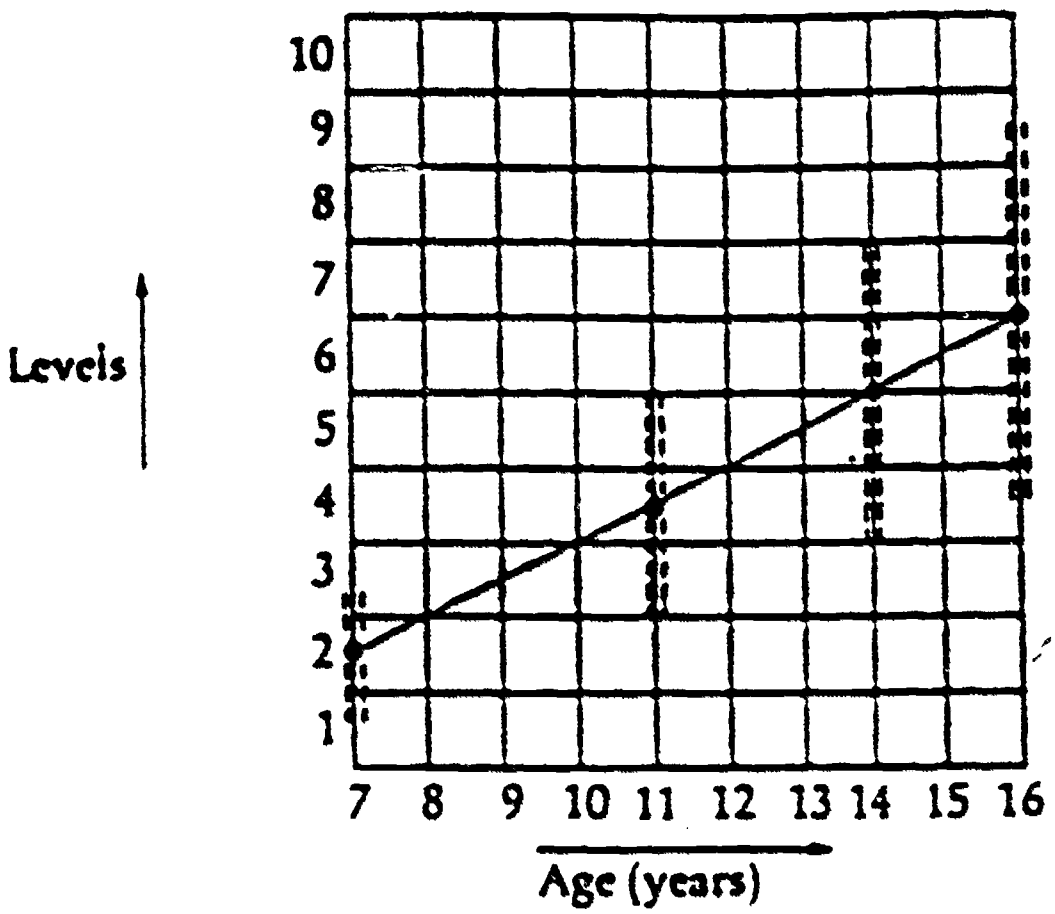
Introduction

In this paper I want to give an account of the use of performance based assessment on a national scale with seven year olds in the UK; to consider the impact of that assessment programme on teaching practice; and to look at some of the implications for this model of assessment.

The national curriculum structure in England and Wales is complex and based on a ten level hierarchial system of progression. Each subject is divided up into a number of attainment targets or strands, the performance required by pupils at various levels within each attainment target is defined by a number of statements of attainment. These statements of attainment form the assessment criteria in the criterion-referenced national assessment system. National assessment has two strands: Teacher Assessment (TA), this is continuous informal formative assessment by teachers against the statements of attainment, and Standard Assessment Tasks (SATs) which are external assessments at ages 7, 11 and 14, given and marked by the teacher (but with some external moderation) in addition to public exams at 16. The national assessment programme started in earnest in the Spring and Summer of 1991 when all seven year olds (year two) were assessed by their teachers and by the external tests or SATs. The SATs are authentic or performance assessments in English, maths and science and cover the first three levels of attainment within the national curriculum system. In 1991 the children were given the SATs by their class teacher in groups of up to four and the exercise required a minimum of 45 hours classroom time for a class of thirty children.

As part of a research project National Assessment in Primary Schools: an evaluation, my colleagues and I are monitoring the implementation of the assessment system, studying teachers' developing practice in assessment, and the interpretation and use of results. Through focussing on teachers' developing assessment practice and the articulation of the national assessment model we aim to extend the theoretical frameworks of assessment. Our fieldwork is based on in-depth work with teachers: through detailed discussion, interview and observation in a range of primary (age 5 to 11) schools, we are building up a database. We have a sample of 32 schools in four very different school districts spread around the country. The districts and the schools are chosen specifically to represent a complete range of settings both physical, geographical, cultural and socio-economic. Thus our schools range from tiny schools in rural areas through large suburban mixed schools to inner city schools with largely bilingual speaking, disadvantaged populations. We visited each school in the Spring Term of 1991 to discuss their experience of teacher assessment which preceded the formal external testing. In the Summer term of 1991 we revisited each school to observe the teachers giving the SATs. We also again interviewed the teachers and the Headteacher. We are currently in the middle of visiting the same schools again for the second year of national assessment: there are however major changes. Teacher assessment does not have to be completed until after the SATs have been done and the SATs

FIGURE 1 SEQUENCE OF PUPIL ACHIEVEMENT OF LEVELS BETWEEN AGES 7 AND 16



The bold line gives the expected results for pupils at the ages specified. The dotted lines represent a rough speculation about the limits within which the great majority of pupils may be found to lie.

themselves have been modified. This paper however focuses on the 1991 SAT exercise. Since this testing represented a rare example of a nationwide assessment of pupils on a criterion-referenced authentic or performance assessment model and does therefore offer clear indications for our understanding of the structure and working of this kind of assessment. A point that I should make here is that in the UK we have always been more involved in open ended assessment than in the USA and we have developed high quality educational assessment of various forms. For example the active, oral and practical work developed by the APU, our version of NAEP, graded tests, GCSE our 16+ school leaving exam with coursework assessment, profiles and Records of Achievement. The SAT was but the logical extension of this trend.

The model for national assessment was produced by the Task Group on Assessment and Testing (the TGAT Report, DES 1988) and relied on teacher assessment as the main assessment device with the SAT used to support and moderate teacher assessment. The SAT as originally conceived in this report consisted of packages of tasks administered through a range of modes including practical, oral, extended and group tasks in order to ensure validity and good curriculum backwash; in addition at primary school level it was considered especially important that the tasks reflect good classroom practice so that the children were not aware that they were being assessed. The 1991 SATs for seven year olds did by and large follow these guidelines. For example, multiplication, subtraction and addition were assessed through children throwing dice as in a game and having to add or multiply the numbers thrown on the dice; floating and sinking in science was assessed through a practical task in which the children were provided with a range of objects and a large tank of water. The children had to predict which objects would float or sink and try and develop a hypothesis (since it could take a week or more to assess a whole class of children on this particular task at one point in the Summer term every infant school classroom could be seen to be full of water, waterlogged objects and rotting pieces of fruit: all the children were reported to have enjoyed it!); at level two reading was assessed by children reading aloud from a book they chose from a range of good children's story books (the list of 20 story books to be used at this level was published first in a national newspaper, within a week all the books were out of stock from bookshops); they were assessed by their teachers for fluency as they read and then asked questions when they had finished reading in order to test their comprehension. In addition there were some paper and pencil tasks to be done in maths on an illustrated work sheet and a story to be written in order to assess writing. In the majority of tasks however the children did not have to write their answers. Teachers were allowed to help the children produce the written answer e.g. in science, and were allowed to make their own judgements about whether the child understood or was able to do the task in hand. Bilingual children were allowed to have an interpreter for the maths and the science tasks. The only attainment target not to be assessed at all by a SAT was listening and speaking: early on in the development process it was decided that this was better assessed by teachers' own judgement.

Findings

Due to the style of assessment, with children having to be assessed individually or in small groups, considerable changes were required to school organisation in order to support the class teachers and to cater for the children who were not being assessed. Schools where team teaching took place and schools where classes were not composed entirely of seven year old pupils generally found this task easier. Schools where there were classes made up entirely of seven year old pupils however had the most re-organisation to effect. By no means all class teachers were offered support to look after the rest of the class while they were assessing children although all teachers were offered some kind of support. Considerable changes were made in some schools to support the administration of the testing. This often had a knock on effect on other staff and where disruption was widespread, e.g. removal of all in-class support from other classes to the year two class, removal of year two teachers from all playground and other duties, it contributed to stress within the school as a whole. However, collegial support for year two teachers was the rule rather than the exception: colleagues offered high levels of support in order to protect the year two teachers from what was seen as an appallingly difficult, stressful and time-consuming activity rather than to perform the assessments particularly well or quickly. In half of our schools the Headteachers were actively involved in doing the SATs or support activities and welcomed the opportunity to spend time with the children. Stress was due not just to the added pressure of having to do the assessment but also to the enormously high level of publicity that the assessments received, hitherto unheard of at primary level, but also to many teachers' anxiety about formally assessing children as young as this with assessments which they felt could be used for labelling children. The culture of our primary teachers maintains that assessment of young children should be only for diagnostic purposes that labelling, and indeed sorting children according to ability, is improper particularly at an age as young as 7 where many children will only have had five full terms of schooling. Teachers are all too well aware of the effect of different lengths of time in school due to birth date, different types of pre-school provision, different family and social backgrounds (especially for ethnic minority children from non-English speaking homes) on children's performance. Thus stress was due to a range of factors related to: a major innovation, the fact of assessment, and the high profile of the activity. An anxiety that we had because of the complexity of the assessment programme and the stress that it caused was that it would succeed in turning teachers off assessment in general; this was a very real danger.

At the time of the testing programme teachers were widely reported as saying that the whole exercise was a waste of time and had told them nothing that they did not already know. We asked our teachers about the usefulness of the SAT experience. Did SATs offer any insights into teaching and learning? Would teachers change their practice in any way? Were there any implications for curriculum development? Out of our 32 schools, teachers at only one stated categorically that they had learned nothing whatsoever; SATs were not revealing anything the teachers did not know already. Fifteen schools (i.e.

teachers and Head) however began by saying that they had learned nothing to further their understanding, but later during the interviews contradicted this. 16 out of the 32 schools felt that administering SATs had been a useful experience, offering room for reflection on curriculum content, individual children's learning, and/or the educational process. One head who carried out an in-house evaluation with her teachers summed up their comments: 'it has certainly taught the staff much about organisation, team-work, forward thinking, planning, assessing, teaching and last but not least enjoyment'. So, virtually all our schools found that they had learned something from the SAT experience. This is in contrast to the widely reported 'they told me nothing I did not already know' comments.

The lessons learned related to curriculum areas on which teachers intended to focus specific attention and resulted in plans to broaden their curriculum. A number of teachers identified gaps in curriculum coverage: they felt they had learned that their children were unpractised in certain specific areas, particularly in English and maths: 21 respondents in 19 schools said that it was their intention to 'do more' in these specific areas next year. Examples of this are giving children more practice in 'instant' multiplication and number bonds, giving children more practice using capital letters and fullstops. In addition to these concerns about specific curriculum content teachers proposed to broaden the curriculum: do applied mathematics more, offer more genres for writing, do more investigational science, review the sort of questions they ask children. Some of these areas for example, instant recall of number bonds, were to be included in order to give the children a fair chance in the assessments despite the teachers' general feeling that it was not appropriate to teach this particular skill at this age. Nevertheless, if it was in the national curriculum for this age and if the children had to be able to do it it was in the children's best interests to make sure that it was covered. Other, for example, investigational science, were to be introduced because the teachers had seen that these activities were worthwhile, within the children's abilities and enjoyable. Other areas which teachers reported intending to review were related to allowing children to do more group work, and encouraging children to work independently, [which is ironic given the current move towards forcing more formal class-based teaching in primary schools.] In some schools, doing the SATs in small groups had meant that teachers were doing group work for the first time, and were encouraged to develop group work: these teachers had discovered that their children could actually work in groups and were gaining a lot from the experience. For some teachers having to leave the rest of the class to get on with independent work was also an eye opener. For these teachers, encouraging independence for the children as learners was shown to be feasible and important. I should point out here that in fact this was the minority of teachers, since at this age most children are taught not in formal class settings with desks facing the front but in a more informal atmosphere with little class teaching.

Whilst the teachers reported that there were lessons to be learned in relation to curriculum coverage, styles of teaching and classroom management they reported that there was little to be learned in relation to assessment per se beyond finding close observation of individual children rewarding ('if only we

had the time...'). There was also little acceptance that the detailed assessment task had shown anything about individual children which the teachers' own informal assessment had not. As researchers we found this last point surprising since for many of these teachers such detailed attention to individual children's performance across such a wide curriculum range was indubitably a new exercise. As for the development of teachers' assessment practice it seems clear to us in this our second year's visits to schools that assessment practice in the schools has changed and developed. There are of course some teachers who have been isolated from the national assessment programme and who still have much to take on board about assessment, but in schools where the assessment tasks were shared, expertise is becoming more widespread.

Standardisation of these assessments was enormously problematic. Instructions to teachers were not specific beyond making certain that the child understood the task. Whilst this is of course entirely appropriate for assessing very young children the lack of standardised introduction for the assessment tasks meant that there was great variation across teachers and also between administrations by the same teacher. We ourselves did not collect data on that particular aspect in 1991. We are doing it rigorously now in 1992. In addition, the statements of attainment are not always sufficiently clear to allow teachers to make ambiguous judgements about performance; the criteria in this criterion-referenced assessment system were not specific enough for assessment purposes. In some schools, which we describe as analytic, teachers discussed criteria and standards of performance among themselves and in these schools it is likely that assessments were more standardised and more comparable across classes than in other schools. In the schools where discussion did take place it was partly because of the woolliness of the assessment criteria that these discussions were started. Teachers' expertise in relation to criterion-referenced assessment is still embroynic: there was evidence that they were reacting to the outcome level in relation to some form of ranking and/or norming for example 'that's not a level three child'. In terms of content validity, it was this very issue that caused some of the manageability problems: because the tasks in many cases matched good infant school practice they were by their very nature time-consuming.

Performance assessments cannot be done in large groups with very young children; in order to deal with the manageability issue the assessments for this summer are less time consuming and less performance-based so that they may be given to whole classes of children. I believe that there are very specific issues related to the age of the children being assessed which means that they require a different format for an assessment programme. For example, our teachers commonly tried to get the best performance out of the children: by reassuring them, helping them, offering preparation and emotional support and sometimes even a second chance. This is one of the criteria for educational assessment (Wood 1986) and definitely runs counter to the notion of assessment as examination or hurdle. This we felt was not due to teachers' particular models of assessment but rather to their view of what is appropriate for children of this age. Teachers were concerned about 'failure' and 'labelling' for such young children and there was some tension between offering children

the chance to try the next advanced level in the assessment programme or indeed to keep plugging away at a particular assessment task, and the need to prevent the children experiencing failure. Our teachers also went to enormous lengths to hide the fact that this was testing; despite the stress and anxiety reported there was very little of this when the children were being assessed. The children were generally unaware of the purpose and importance of the tasks that they were engaged in. This was because the teachers were at great pains to ensure that they were protected from what was going on. Very few children were seen to be upset by the activities, some were bored but it was much more common that children enjoyed them.

It may also be the case that when teachers of young children assess those children, either individually or in small groups, it is almost inevitable that they will vary the way in which they introduce the task whether they are given highly specific instructions or general instructions; this is because what the teacher sees is not a testing situation but individual children whom she or he knows well and who need to have things explained to them in different ways, or presented in different ways because of the children's own backgrounds, abilities and immediate past history. If this is the case then it is not possible (and one might say not desirable) to have highly standardised performance assessments, because performance assessments with young children must be done in small groups.

Implications

I have described our early observations of the introduction of a high stakes national performance assessment of young children. There is no doubt that the SATs represent authentic/performance assessment and by and large they matched the active process-based tasks which children do in good infant classroom practice. As our data shows, these assessment tasks not only gave our teachers direct feedback about areas of the curriculum which they had not covered, but also pointers towards a wider view of teaching and learning. This is the opposite of the traditional concept of teaching to the test, which is typically viewed as narrowing and negative, in that it widened some teachers' practice rather than narrowed it. I believe we have shown, albeit in a small way, that high stakes, performance assessments can improve the teaching of higher order skills (Shepard 1991). Unfortunately, we have a second opportunity to observe changes in teachers' practice, possibly back towards a narrowing again, as this year's SATs are less active and less process-based.

There are of course serious issues in relation to validity and reliability in the SAT assessment. A number of these I have already commented on. In addition to the unreliability resulting from differing administration styles by teachers and different interpretations of children's performance there were major areas of unreliability in the reading test. The reading test which involved children reading aloud from real books and then being asked questions about the content and future events was high on content validity in that it matches what we think of as real reading for average seven year olds. However, part of the attempt to enhance validity was a very cause of the unreliability: there was a choice allowed from a range of 20 books and it was not

uncommon for children to be reading from a book which they, it turned out, knew well. Thus obviously for some children the task was much easier since they already knew the story and had practised the reading. Our view is that it would have been better to produce a range of specially written books for this exercise which look like good quality children's literature but would be unknown to all children. This has still not happened. It is of course the balance between reliability and validity which is the nub of the problem. One of the criticisms of performance assessment that is coming through to us from the American literature is that performance assessments can be just as unreliable and invalid as traditional tests, potentially more so because they rely on fewer tasks. Performance assessment can only cover a limited range of activities, particularly with younger children because they are by nature time consuming and may need to be carried out individually or in small groups, but in the UK this should not be a problem because we have the requirement for teacher assessment as well. In other words what we would be looking for is a combination of high quality, time consuming performance assessment which covers a changing, smaller number of skills complimented by teacher assessment of a much wider range of skills. The advantage of the high quality performance assessment here is that over the years it can come to support and moderate teacher's own assessment practice.

Now I want to look at the national assessment programme in relation to some of the validation criteria put forward for performance based assessment by Linn, Baker and Dunbar (Linn et al, 1991). I have already looked at the issue of consequences of the SATs for teachers of seven year olds; for some teachers the SATs did serve as examplars of good teaching practice which widened teachers' models of curriculum and teaching style. We do not know however how widespread such an effect was, but we intend to replicate our study this year when the teachers have to conduct another national survey but on a different model of SAT.

From other work on the national assessment programme (see Gipps C 1992) we know that the SATs were seen by many teachers as a more fair way of assessing bilingual children, and children with special educational needs, than group standardised paper and pencil tests would be. Despite their heavy reliance on language, teachers felt that with the interactive nature of the assessment, such pupils were given a fair chance to show what they knew and understood with the result (incomprehensible to some on the right wing) that children who can barely write were judged as being at level one or even level two in science i.e. had a working knowledge of some basic scientific concepts. In the piloting of SATs for fourteen year olds in Summer 1991 comments about the maths tasks match general comments about the SATs for seven year olds. The teachers of fourteen year olds who were not fluent in English regarded the nature of the SAT as rendering it accessible to pupils who were not fluent in the language. It was a combination of interaction with the teacher, the practical elements of the task, a normal classroom atmosphere, interactions with other pupils and a variety of presentation and assessment modes which contributed to this. Teachers of such pupils felt that written materials alone could not allow the demonstration of potential and understanding without teacher-pupil interaction. "If pupils who are not fluent in English are to be entitled to a fair

assessment it is essential that the SATs retain the interactive, practical and flexible aspects". (CATS 1991).

The quality of the content and the tasks themselves was felt generally to be adequate, if not in some cases good, but certainly better than traditional examinations or standardised tests would be able to offer. The same is true for cognitive complexity, indeed, there was evidence that in a number of cases teachers had to revise upwards their views about children's attainment (both at 7 and 14) because they were obliged to allow them to do certain tasks in the SATs which they had thought the pupils would not be capable of doing. The tasks were on the whole meaningful and worthwhile educational experiences teachers felt and what is more they were enjoyable; the teachers had no concerns in that department. The main problem which proved to be their downfall was the non-manageability: the assessment was time-consuming and, following any definition, unmanageable; this model of SAT was simply not appropriate for large-scale assessments. What we have not yet been able to do is to study in detail the use of results though we have this in hand. We do not yet know the extent to which instruction for individual pupils is altered as a result of the assessment results, or the extent to which national assessment serves its accountability function. It is not yet clear to us how parents have understood the results nor how they would wish the school to act on them. These features are, however, part of our ongoing research project and we hope to be able to report on them next year.

<u>What went wrong?</u>

The original model for national assessment relied on teacher assessment as the main assessment device with the SAT used to support and moderate teacher assessment. However, many readers of the TGAT Report understood that each SAT was to be given to each pupil and the SAT result weighed up with the teacher assessment result to give a final figure for each child and this was the basis for the development of the SAT programme. If the SATs are used to assess individual children they must therefore produce results which are reliable as well as valid at the individual level. Reliability is of course highly significant if assessment results are to be used to make comparisons of pupils and schools and this was always a prime purpose for our national assessment programme. There is however another reading of the TGAT Report: that the SATs were to be used, not to provide results for individual children, but to moderate a teacher's overall results for a class. This of course is a highly significant difference in interpretation, and the authors of the report say that their intended model was the latter. If the SAT is only an overall moderating device, then it need only sample across the curriculum. If it has only to sample across the curriculum, and indeed across children, then detailed time-consuming assessments are possible. If however, SATs are to be used to confirm teacher assessment for each child then they must cover each element of the curriculum. The SAT model as originally developed is not appropriate for assessing literally hundreds of assessment points on whole age groups of children at any one point in time: for what is essentially survey testing, something quicker and more reliable is needed. The original SAT model is ideal for individual assessment by teachers for formative and diagnostic

purposes. This individual teacher-based assessment can of course be summed up at the end of each stage of education to give summative information. However, in the UK we tend to take the view that summative assessment, particularly if it is also to be used for evaluative purposes or for certification and selection, must be taken out of teachers' hands. Thus, teacher assessment is not to be used at the end of the key stages in education (i.e. 7, 11, 14 and 16) because teacher assessment is liable to be unreliable and/or biased. It is of course true that teachers do need some form of referencing if their standards are to be comparable across the country which fairness and equity demand. For our public exams at 16 and 18 a combination of external marking and moderation processes have been developed to deal with this issue and it is widely accepted that this produces reliable judgements (though this is questionable). However, an assessment system which relied on widespread moderation and extended marking when applied to four age groups simultaneously is clearly unmanageable.

In the TGAT Report, the emphasis was on the professional uses of assessment. There was little mention of standards and accountability procedures. The tone of the report was thus at odds with the political climate within which national curriculum and assessment was introduced. Small wonder then that as teachers complained of the workload involved in SATs, and the low level of standardisation became clear, the Prime Minister said that the SATs for 1992 would be largely paper and pencil tests, standardised, and capable of being taken by the whole class at once. The formal, unseen, examination had served the system well in the past and would do so again. As Linn et al put it "... if great weight is attached to the traditional criteria of efficiency, reliability, and comparability of assessments from year to year, the more complex and time-consuming performance-based measures will compare unfavourably with traditional standardised tests." (Linn et al 1991, op cit) Another problem with the original TGAT model was that it suggested that the same system of assessment could serve all required purposes: formative, diagnostic, summative and evaluative. The notion that one programme of assesment could fulfill four functions was always questionable and has been shown to be false: different purposes require different models of assessment and different relationships between teacher and pupil. It may be possible to design one assessment system which measures performance at school level for accountability purposes and at individual pupil level for selection purposes whilst at the same time supporting the teaching-learning process but we have not yet done it. Assessment for formative purposes is essentially carried out by the teacher in an informal way, often with no clear conclusions, but the repeated assessment at an informal level allows the teacher to form valid assessments of the pupil's performance particularly because s/he is able to assess the pupil in a number of settings and contexts. External assessment for summative and evaluative purposes tends to be one-off and external to the teacher-pupil relationship. The final straw was the complexity of the curriculum structure: this resulted in an enormous number of statements of attainment which became the assessment criteria in the criterion-referenced assessment system. Requiring teachers to assess every child on every criterion and to report this four times during their school career is difficult enough, but to link this with external, project-type assessment of every pupil on a high

proportion of these criteria at a particular point in the school year is clearly too daunting and time-consuming a task.

There is a lot to be learned from our attempts to develop the national assessment system. Much of it as yet untapped. My colleagues and I have in another setting articulated an alternative model for a national assessment programme (BERA 1992). In the meantime, we must continue our attempts to develop high quality process-based assessment which can support teaching and learning and which, if it is to be used for accountability and comparison purposes, is able to optimise reliability without jeopardising validity. The problems that we had in the UK arising from the complexity of the underlying curriculum structure, the inappropriateness of the model for what was in effect a national survey, and a harsh political climate, should not divert us from the search for good quality educational assessment. Nor should the problems of development divert us from the message that good quality educational assessment, although time and resource demanding, is the best way forward.

Acknowledgements

This research is supported by the Economic and Social Research Council grant number 000 23 2192.

proportion of these criteria at a particular point in the school year is clearly too daunting and time-consuming a task.

There is a lot to be learned from our attempts to develop the national assessment system. Much of it as yet untapped. My colleagues and I have in another setting articulated an alternative model for a national assessment programme (BERA 1992). In the meantime, we must continue our attempts to develop high quality process-based assessment which can support teaching and learning and which, if it is to be used for accountability and comparison purposes, is able to optimise reliability without jeopardising validity. The problems that we had in the UK arising from the complexity of the underlying curriculum structure, the inappropriateness of the model for what was in effect a national survey, and a harsh political climate, should not divert us from the search for good quality educational assessment. Nor should the problems of development divert us from the message that good quality educational assessment, although time and resource demanding, is the best way forward.

Acknowledgements

This research is supported by the Economic and Social Research Council grant number 000 23 2192.