Cambridge Assessment

Cambridge Assessment is Europe's largest assessment agency and plays a leading role in researching, developing and delivering assessment across the globe. It is a department of the University of Cambridge and a not-for-profit organisation with a turnover of around £175 million. The Group employs around 1,400 people and contracts some 15,000 examiners each year.

Cambridge Assessment's portfolio of activities includes world-leading research, ground-breaking new developments and career enhancement for assessment professionals. Public examinations and tests are delivered around the globe through our three highly respected examining bodies.

The assessment providers in the Group include:

University of Cambridge English for Speakers of Other Languages (Cambridge ESOL)
Tests and qualifications from Cambridge ESOL are taken by over 1.75 million people, in 135 countries. Cambridge ESOL's Teaching Awards provide a route into the English Language Teaching profession for new teachers and first class career development opportunities for experienced teachers. Cambridge ESOL works with a number of governments in the field of language and immigration.

University of Cambridge International Examinations (CIE)
CIE is the world's largest provider of international qualifications for 14-19 year-olds. CIE qualifications are available in over 150 countries. CIE works directly with a number of governments to provide qualifications, training and system renewal.

OCR
The Oxford Cambridge and RSA awarding body (OCR) provides general academic qualifications and vocational qualifications for learners of all ages through 13,000 schools, colleges and other institutions. It is one of the three main awarding bodies for school qualifications in England.

ARD
The Assessment Research and Development division (ARD) supports development and evaluation work across the Cambridge Assessment group and administers a range of admissions tests for entry to Higher Education. The ARD includes the Psychometrics Centre, a provider and developer of psychometric tests.

1. A view of the scope of the enquiry

1. At Cambridge Assessment we recognize that it is vital not to approach assessment on a piecemeal basis. The education system is exactly that: a system. Later experiences of learners are conditioned by earlier ones; different elements of the system may be experienced by learners as contrasting and contradictory; discontinuities between elements in the system (e.g. transition from primary to secondary education) may be very challenging to learners.

2. Whilst understanding the system as a system is important, we believe that the current focus on 14-19 developments (particularly the development of the Diplomas and post-Tomlinson strategy) can all too readily take attention away from the serious problems which are present in 5-14 national assessment.

3. Our evidence tends to focus on assessment issues. This is central to our organisation's functions and expertise. However, we are most anxious to ensure that assessment serves key functions in terms of supporting effective learning (formative functions) and progression (summative functions). Both should be supported by effective assessment.

4. We welcome the framing of the Committee's terms of reference for this Inquiry, which make it clear that it intends to treat these two areas as substantially discrete. Cambridge Assessment's qualifications deliverers (exam boards), OCR and University of Cambridge International Examinations, have tendered evidence separately to this submission. They have looked chiefly at 14-19 qualifications.

5. This particular submission falls into two sections: Firstly Cambridge Assessment's views on the national assessment framework (for children aged 5-14). These are informed by, but not necessarily limited to, the work which we carried out through out 2006 in partnership with the Institute for Public Policy Research (IPPR) and the substantial expertise in the Group of those who have worked on national tests.

6. The second section is on University Entrance Tests. Cambridge Assessment has been involved in the development of these for nearly a decade and uses a research base that stretches back even further. At first their scope was limited to Cambridge University but over the last four years it has grown to include many other institutions. That they are administered under Cambridge Assessment's auspices (as opposed to those of one of our exam boards) is a reflection of their roots within our research faculty and the non statutory nature of the tests themselves.

Section 1
2. National assessment arrangements

7. In this section we have sought to outline the problems that have built up around the national assessment arrangements. We have then gone on to discuss the changes proposed in our work with the IPPR. We also then discuss the problems that appear to be inherent in the 'Making Progress' model that the Government is committed to trialling. Our conclusion is that there is a window of opportunity before us at the present time, just one of the reasons that the Committee's Inquiry is so timely, and that the Government should not close it with the dangerous haste that it seems bent on. There are a range of options and to pursue only one is a serious mistake.

8. We have included two Annexes:

- an overview of the evidence on national assessment dealing with questions ranging from **'**teaching to the test' to 'measurement error'
- a brief discussion of why the sometimes mooted return of the APU might not deliver all the objectives desired of it.

3. Diagnosis of the Challenge – critique and revision of national assessment arrangements

9. It is important to note that Cambridge Assessment is highly supportive of the principle of a National Curriculum and related national assessment. The concept of 'entitlement' at the heart of the National Curriculum has been vital to raising achievement overall; raising the attainment of specific groups (e.g. females in respect of maths and science); and ensuring breadth and balance. We recognise that enormous effort has been put in, by officials and developers, to improving the tests in successive years. We support the general sentiment of the Rose Review – that the system has some strong characteristics – but it is clear that deep structural problems have built up over time.

10. Whilst being concerned over these problems, Cambridge Assessment is committed to the key functions supported by national assessment: provision of information for formative and diagnostic purposes to pupils, teachers and parents; information on national standards, and accountability at school level. We return to these key functions in more detail below. However, Cambridge Assessment is critical of the way in which national assessment has been progressively and successively elaborated into a system which appears to be yielding too many serious and systemic problems.

Accumulating problems in National Assessment – a vessel full to bursting point?

11. There are two particularly significant problems in the highly sensitive area of technical development of national assessment arrangements. Firstly, previous statements by agencies, departments and Government have exaggerated the technical rigour of national assessment. Thus any attempts to more accurately describe its technical character run the risk of undermining both the departments and ministers; '…if you're saying this now, how it is that you said that, two years ago…'. This prevents rational debate of problems and scientifically-founded development of arrangements.

Secondly, as each critique has become public, the tendency is to breathe a sigh of relief as the press storm abates; each report is literally or metaphorically placed in a locked cupboard and forgotten.

12. In contrast, we have attempted here to take all relevant evidence and integrate it; synthesising it in such a way that underlying problems and tendencies can accurately be appraised – with the intention of ensuring effective evaluation and refinement of systems.

14. Put simply, if a minister asks a sensible question: '…are attainment standards in English going up or down and by how much?...' there is no means of delivering a valid and sound response to that question using current arrangements. This is a serious problem for policy formation and system management. It is not a position which obtains in systems which use independent light sampling methods such as the US NAEP (National Assessment of Educational Progress).

Functions

15. Current national curriculum assessment arrangements within England have attracted increasing criticism in respect of the extent to which they are carrying too many purposes (Brooks R & Tough S; Bell J et al; Daugherty R et al). Since 1988 a substantial set of overt and tacit functions have found themselves added. The original purposes specified in the TGAT Report (Task Group on Assessment and Testing) comprised:

1 formative (diagnostic for pupils; diagnostic for teachers)
2 summative (feedback for pupils and parents)
3 evaluative (providing information at LEA and school level)
4 informative (providing information on educational standards at system level)

16. The following have been added, as increasingly elaborated uses of the flow of detailed data from national assessment:

- school accountability
- departmental accountability
- apportionment of funds
- inspection patterns and actions
- upwards pressure on standards/target setting
- structuring of educational markets and school choice
- emphasis of specific curriculum elements and approaches
- detailed tracking of individual attainment, strengths and weaknesses
- quantification of progress

17. Unsurprisingly, many educationalists have expressed the view that the current tests carry too many functions and that the underlying management processes are too elaborated. To carry this broad range of functions, the system of assessing every child at the end of each Key Stage is dependent on maintaining test standards over time in a way which is in fact not practical.

18. If you want to measure change, don't change the measure. But the nation does – and should – change/update the National Curriculum regularly. Whenever there is change (and sometimes radical overhaul) the maintenance of test standards becomes a particularly aggressive problem. It does, of course, remain a constant problem in areas such as English Literature when one could be pretesting a test on *Macbeth* which will be taken in 2008 but the pupils are currently studying *As You Like it* when they sit the pretest. There are remedies to some of the problems this creates – namely switch to different sampling processes; announcing radical recalibration, or switch to low stakes sampling of children's performance, using a NAEP or a modernized APU-style model (Assessment of Performance Unit – see Annexe 2)

19. Attempting to use national assessment to measure trends over time has produced some of the most intense tensions amongst the set of functions now attached to national testing. Stability in the instruments is one of the strongest recommendations emerging from projects designed to monitor standards over time. Running counter to this, QCA and the DfES have – in line with commitments to high quality educational provision, the standards agenda and responses from review and evaluation processes - sought to optimize the National Curriculum by successive revision of content, increasing the 'accessibility of tests', and ensuring tight linkage of the tests to specific curriculum content.

20. These are laudable aims – and the emphasis on the diagnostic function of the data from tests has been increasing in recent innovations in testing arrangements However, pursuit of these aims has led to repeated revision rather than stability in the tests. The Massey Report suggested that if maintenance of standards over time remained a key operational aim, then stability in the test content was imperative (Massey A et al). In the face of these tensions, a light sampling survey method would enable de-coupling of national assessment from a requirement to deliver robust information on national educational standards. This would enable testing to reflect curriculum change with precision, to optimize the learning-focussed functions of testing, and enable constant innovation in the form of tests to optimize accessibility.

21. It is therefore clear that the current functions of national testing arrangements are in acute and chronic tension. Using the pragmatic argument that 'every policy should have a policy instrument' we conclude that national arrangements should indeed support school accountability and improvement, report to parents and monitor national standards but that a change of arrangements is required to achieve this. A range of approaches are necessary to deliver these functions and we outline some viable options below.

3. Alternative Approaches to National Assessment (KS1, KS2, KS3)

Objectives

22. There is a need to conceptualise a number of possible models for consideration in an attempt to address the problems of 'multipurpose testing'. It is vital to note that we present here three alternatives. We do this to show that there are credible alternatives for delivering on the key objectives of national assessment – it is simply not the case that there is only one way of moving forward.

23. We believe the aims should be to

- reduce the assessment burden on schools
- provide formative assessment for teaching and learning
- provide information for school accountability
- provide information on national standards.

24. In order to secure widespread support within the education community (including parents) a firm re-statement of educational purpose (values) and a commitment to high degrees of validity is essential. It is not enough to initiate changes merely because of concerns about the defects of existing arrangements. We do not here outline values and validity in detail, but recognise that this is a vital precondition of designing revised arrangements, putting them in place, and monitoring their operation. It is important that a full discussion of these matters precedes any executive decision regarding revised arrangements.

Alternative models for national assessment

Model 1: Validity in monitoring plus accountability to school level

25. The aim of this approach is to collect data using a national monitoring survey and to use this data for monitoring standards over time as well as for moderation of teacher assessment. This would enable school performance to be measured for accountability purposes and would involve a special kind of criterion referencing known as domain referencing.

26. Question banks would be created based on the curriculum with each measure focusing on a defined domain. A sample of questions would be taken from the bank and divided into lots of small testlets (smaller than the current KS tests). These would then be randomly allocated to each candidate in a school. Every question is therefore attempted by thousands of candidates so the summary statistics are very accurate and there are summary statistics on a large sample of questions. This means that for a particular year it is known, for example, that on average candidates can obtain 50% of the marks in domain Y.

27. The following year it might be found that they obtain 55% of the marks in that domain. This therefore measures the change and no judgement about relative year-on-year test difficulty is required. Neither is there a need for a complex statistical model for analysing the data, although modelling would be required to calculate the standard errors of the statistics reported. However, with the correct design they would be superfluous because they would be negligible. It would be possible to use a preliminary survey to link domains to existing levels and the issue of changing items over time could be solved by chaining and making comparisons based on common items between years. Although each testlet would be an unreliable measure in itself, it would be possible to assign levels to marks using a statistical method once an overall analysis had been carried out. The average of the testlet scores would be a good measure of a school's performance given that there are sufficient candidates in the school. The appropriate number of candidates would need to be investigated.

28. The survey data could also be used to moderate teacher assessment by asking the teacher to rank order the candidates and to assign a level to each of them. Teacher assessment levels would then be compared with testlet levels and the differences calculated. It would not be expected that the differences should be zero, but rather that the need for moderation should be determined by whether the differences cancel out or not. Work would need to be done to establish the levels of tolerance and the rules for applying this process would need to be agreed. The school could have the option of accepting the statistical moderation or going through a more formal moderation process.

29. There would be a number of potential advantages related to this model. Validity would be increased as there would be greater curriculum coverage. The data would be more appropriate for the investigation of standards over time. The test development process would be less expensive as items could be re-used through an item bank, including past items from national curriculum tests. There would also be fewer problems with security related to 'whole tests'. No awarding meetings would be needed as the outcomes would be automatic and not judgemental. Since candidates would not be able to prepare for a specific paper the negative wash-back and narrowing of the curriculum would be eliminated (i.e. the potential elimination of 'teaching to the test').. There would also be less pressure on the individual student since the tests would be low stakes.

30. Given that there are enough students in a school, the differences in question difficulty and pupil question interaction would average out to zero leaving only the mean of the pupil effects. From the data it would be possible to generate a range of reports e.g. equipercentiles and domain profiles. Reporting of domain profiles would address an issue raised by Tymms (2004) that 'the official results deal with whole areas of the curriculum but the data suggests that standards have changed differently in different sub-areas'.

31. Work would need to be done to overcome a number of potential disadvantages of the model. Transparency and perception would be important and stakeholders would need to be able to understand the model sufficiently to have confidence in the outcomes. This would be a particularly sensitive issue as students could be expected to take tests that prove to be too difficult or too easy for them. Some stratification of the tests according to difficulty and ability would alleviate this problem. There is an assumption that teachers can rank order students (Lamming D) and this would need to be explored. Applying the model to English would need further thought in order to accommodate the variations in task type and skills assessed that arise in that subject area.

32. Eventually the model would offer the possibility of reducing the assessment burden but the burden would be comparatively greater for the primary phase. Although

security problems could be alleviated by using item banking, the impact of item re-use would need to be considered. Having items in the public domain would be a novel situation for almost any other important test in the UK (except the driving test).

33. Discussion and research would be needed in a number of areas

- values and validity
- scale and scope e.g. number and age of candidates, regularity and timing of tests
- formal development of the statistics model
- simulation of data (based on APU science data initially)
- stratification of tests / students
- pilots and trials of any proposed system

Model 2: Validity in monitoring plus a switch to 'school-improvement inspection'

34. Whilst the processes for equating standards over time have been enhanced since the production of the Massey Report, there remain significant issues relating to:

- teacher confidence in test outcomes
- evidence of negative wash-back into learning approaches
- over-interpretation of data at pupil group level; inferences of improvement or deterioration of performance not being robust due to small group size
- ambiguity in policy regarding borderlining
- no provision to implement Massey recommendations regarding keeping tests stable for 5 years and then 'recalibrating' national standards
- publishing error figures for national tests

35. In the face of these problems, it is attractive to adopt a low-stakes, matrix-based, light sampling survey of schools and pupils in order to offer intelligence to Government on underlying educational standards. With a matrix model underpinning the sampling frame, far wider coverage of the curriculum can be offered than with current national testing arrangements.

36. However, if used as a replacement for national testing of every child at the end of KS1, 2 and 3, then key functions of the existing system would not be delivered:

- data reporting, to parents, progress for every child at the end of each key stage
- school accountability measures

37. In a system with a light sampling model for monitoring national standards, the first of these functions could be delivered through (i) moderated teacher assessment, combined with (ii) internal testing, or tests provided by external agencies and/or grouped schools arrangements. The DfES prototype work on assessment for learning could form national guidelines for (i) the overall purpose and framework for school assessment, and (ii) model processes. This framework of assessment policy would be central to the inspection framework used in school inspection.

38. The intention would be to give sensitive feedback to learners and parents, with the prime function of highlighting to parents how best to support their child's learning. Moderated teacher assessment has been proven to facilitate staff development and

effective pedagogic practice. Arrangements could operate on a local or regional level, allowing transfer of practice from school to school.

39. The second of these functions could be delivered through a change in the Ofsted inspection model. A new framework would be required since the current framework is heavily dependent on national test data, with all the attendant problems of the error in the data and instability of standards over time. Inspection could operate through a new balance of regional/area inspection services and national inspection – inspection teams operating on a regional/area basis could be designated as 'school improvement teams'. To avoid competition between national and regional inspection, national inspections would be joint activities led by the national inspection service.

40. These revised arrangements would lead to increased frequency of inspection (including short-notice inspection) for individual schools and increased emphasis on advice and support to schools in respect of development and curriculum innovation. Inspection would continue to focus on creating high expectations, meeting learner needs, and ensuring progression and development.

Model 3: Adaptive, on-demand testing using IT- based tests

41. In 2002, Bennett outlined a new world of adaptive, on-demand tests which could be delivered through machines. He suggests that 'the incorporation of technology into assessment is inevitable because, as technology becomes intertwined with what and how students learn, the means we use to document achievement must keep pace'. Bennett (2001) identifies a challenge, 'to figure out how to design and deliver embedded assessment that provides instructional support and that globally summarises learning accomplishment'. He is optimistic that 'as we move assessment closer to instruction, we should eventually be able to adapt to the interests of the learner and to the particular strengths and weaknesses evident at any particular juncture…'. This is aligned to the commitments of Government to encourage rates of progression based on individual attainment and pace of learning rather than age-related testing.

42. In the Government's five year strategy for education and children's services (DfES, 2004) principles for reform included 'personalisation and choice as well as flexibility and independence'. The White Paper on 14 – 19 Education and Skills (2005) stated, 'Our intention is to create an education system tailored to the needs of the individual pupil, in which young people are stretched to achieve, are more able to take qualifications as soon as they are ready, rather than at fixed times…' and 'to provide a tailored programme for each young person and intensive personal guidance and support'. These intentions are equally important in the context of national testing systems.

43. The process relies on item-banking, combining items in individual test sessions to feed to students a set of questions appropriate to their stage of learning and to their individual level of attainment. Frequent low-stakes assessments could allow coverage of the curriculum over a school year. Partial repetition in tests, whilst they are 'homing in' on an appropriate testing level, would be useful as a means of checking the extent to which pupils have really mastered and retained knowledge and understanding.

44. Pupils would be awarded a level at the end of each key stage based on performance on groups of questions to which a level has been assigned. More advantageously, levels could be awarded in the middle of the key stage as in the revised Welsh national assessment arrangements.

45. Since tests are individualised, adaptivity helps with security, with manageability, and with reducing the 'stakes', moving away from large groups of students taking a test on a single occasion. Cloned items further help security. This is where an item on a topic can include different number values on a set of variables, allowing the same basic question to be systematically changed on different test administrations, thus preventing memorisation of responses. A simple example of cloning is where a calculation using ratio can use a 3:2 ratio in one item version and 5:3 ratio in another. The calibration of the bank would be crucial with item parameters carefully set and research to ensure that cloning does not lead to significant variations in item difficulty.

46. Reporting on national standards for policy purposes could be delivered through periodic reporting of groups of cognate items. As pupils nationally take the tests and when a critical nationally representative sample on a test is reached, this would be lodged as the national report of standards in a given area. This would involve grouping key items in the bank e.g. on understanding 2D representation of 3D objects and accumulating pupils' performance data on an annual basis (or more or less frequently, as deemed appropriate) and reporting on the basis of key elements of maths, English etc.

47. This 'cognate grouping' approach would tend to reduce the stakes of national assessment, thus gauging more accurately underlying national standards of attainment. This would alleviate the problem identified by Tymms (2004) that 'the test data are used in a very high-stakes fashion and the pressure created makes it hard to interpret that data. Teaching test technique must surely have contributed to some of the rise, as must teaching to the test'.

48. Data could be linked to other cognate groupings, e.g. those who are good at X are also good at Y and poor on Z. Also, performance could be linked across subjects.

49. There are issues of reductivism in this model as there could be a danger to validity and curriculum coverage as a result of moving to test forms which are 'bankable', work on-screen and are machine-markable. Using the Cambridge taxonomy of assessment items is one means of monitoring intended and unintended drift. It is certainly not the case that these testing technologies can only utilise the most simple multiple-choice (mc) items. MC items are used as part of high-level professional assessment e.g. in the medical and finance arenas, where well-designed items can be used for assessing how learners integrate knowledge to solve complex problems.

50. However, it is certainly true that, at the current stage of development, this type of approach to delivering assessment cannot handle the full range of items which are currently used in national testing and national qualifications. The limitation on the range of item types means that this form of testing is best used as a component in a national assessment model, and not the sole vehicle for all functions in the system.

51. School accountability could be delivered through this system using either (i) a school accumulation model, where the school automatically accumulates performance data from the adaptive tests in a school data record which is submitted automatically when the sample level reaches an appropriate level in each or all key subject areas, or (ii) the school improvement model outlined in model 2 above.

52. There are significant problems of capacity and readiness in schools, as evidenced through the problems being encountered by the KS3 ICT test project which has successively failed to meet take-up targets. It remains to be seen whether these can be swiftly overcome or are structural problems e.g. schools adopting very different IT network solutions and arranging IT in inflexible ways. However, it is very important to note that current arrangements remain based on 'test sessions' of large groups of pupils, rather than true on-demand, adaptive tests. These arrangements could relieve greatly the pressures on infrastructure in schools, since sessions would be arranged for individuals or small groups on a 'when ready' basis.

53. There are technical issues of validity and comparability to be considered. The facility of a test is more than the sum of the facility on the individual items which make up each test. However, this is an area of intense technical development in the assessment community, with new understanding and theorisations of assessment emerging rapidly.

54. There are issues of pedagogy. Can schools and teachers actually manage a process where children progress at different rates based on on-demand testing? How do learners and teachers judge when a child is ready? Will the model lead to higher expectations for all students, or self-fulfilling patterns of poor performance amongst some student groups? These – and many more important questions – indicate that the assessment model should be tied to appropriate learning and management strategies, and is thus not neutral technology, independent of learning.

Overall

55. Each of the models addresses the difficulties of multipurpose testing. However, each model also presents challenges to be considered and overcome. The Statistics Commission (2005) commented that 'there is no real alternative at present to using statutory tests for setting targets for aggregate standards'. The task is to find such an alternative. The real challenge is to provide school accountability data without contaminating the process of gathering data on national standards and individual student performance. All three models have their advantages and could lead to increased validity and reliability in national assessment arrangements and – crucially – the flow of reliable information on underlying educational standards; something which is seriously compromised in current arrangements.

4. New progress tests – serious technical problems

56. As a possible line of development for new arrangements, the DfES recently has announced pilots of new test arrangements, to be trialled in 10 authorities. Cambridge Assessment has reviewed the proposals and, along with many others in the assessment community, consider that the design is seriously flawed. The deficiencies are significant enough to compromise the new model's capacity to deliver on the key functions of national assessment; i.e. information on attainment standards at system level; feedback to parents, pupils and teacher; and provision of school accountability.

57. Cambridge Assessment's response to the DfES consultation document on the progress tests covered the subject in some detail and we reproduce it below for the Select Committee.

i
We welcome the developing debate on the function and utility of national assessment arrangements. We applaud the focus on development of arrangements which best support the wide range of learning and assessment needs amongst those in compulsory schooling.

ii
As specialists in assessment, we have focused our comments on the technical issues associated with the proposals on testing. However, it is vital to note that Cambridge Assessment considers fitness for purpose and a beneficial linkage between learning and assessment to be at the heart of sound assessment practice.

iii
We consider effective piloting, with adequate ethical safeguards for participants, to be essential to design and implementation of high quality assessment arrangements. It is essential that evaluation method, time-frames, and steering and reporting arrangements all enable the outcomes of piloting to be fed into operational systems. There is inadequate detail in the document to determine whether appropriate arrangements are in place.

iv
We remain concerned over conflicting public statements regarding the possible status of the new tests (TES march 30th), which make it very unclear as to whether existing testing arrangements will co-exist alongside new arrangements, or whether one will be replaced by the other. This level of confusion is not helpful.

v
We see three functions as being essential to national assessment arrangements:

Intelligence on national standards – for the policy process
Information on individual pupil performance – for the learner, for parents, for teachers
Data on school performance – for accountability arrangements

We do not feel that the new model will meet these as effectively as other possible models. We would welcome discussions on alternatives.

vi

We believe that, by themselves, the new test arrangements will not provide robust information on underlying standards in the education system. With entry to single-level tests dependent on teachers' decisions, teachers in different institutions and at different times are likely to deploy different approaches to entry. This is likely to be very volatile, and effects are unlikely to always cancel out. This is likely to contaminate the national data in a very new ways, compared with existing testing arrangements. There are no obvious remedies to this problem within the proposed arrangements, either in the form of guidance or regulation.

vii
Teachers are likely to come under peculiar pressures, from institutions wishing to optimise performance-table position, from parents of individual children etc. This is an entirely different scenario to the 'all to be tested and then a level emerges' character of current arrangements. Tiering invokes a similar, though not as here all-pervasive, effect.

viii
Although advanced as 'on-demand' testing, the regime is not an 'on-demand' regime, and it is misleading to promote it as such. It provides one extra test session per year.

ix
The frequency of testing is likely to increase the extent to which testing dominates teaching time. This is not a problem where the majority of washback effects from testing are demonstrably beneficial; we believe that other features of the tests mean that washback effects are likely to be detrimental. It is not clear what kind of differentiation in teaching will flow back from the tests. Ofsted and other research shows differentiation to be one of the least developed areas of teaching practices. We are concerned that the 'grade D' problem (neglect of this those not capable of getting a C and those who will certainly gain a C) will emerge in a very complex form in the new arrangements.

x
The tests may become MORE high stakes for learners. Labelling such as '…you're doing level 2 for the third time!...' may emerge and be very pernicious. Jean Rudduck's work shows such labelling to be endemic and problematic.

xi
We are unclear regarding the impact on those learners who fail a test by a small margin – they will wait 6 months to be re-tested. Do teachers judge that they should 'lose 6 months of teaching time' to get them up to the required level or just carry on with no special support. If special support is given, what is the child not doing which they previously would have done? This is a key issue with groups such as less able boys – they will need to take time out of things which they are good at and which can bolster their 'learning identities'. Those who are a 'near miss' will need to know – the document does not make clear whether learners will just 'get a level back'; will get a mark; or an item-performance breakdown.

xii
Testing arrangements are likely to become much more complex – security issues, mistakes (such as wrong test for a child) etc are likely to gain in significance.

xiii

The length of the tests *may* be an improvement over existing tests, but further investigative work must be done to establish whether this is indeed the case. 45-minute tests may, or may sample more from each subject domain at an appropriate level, compared with existing national tests. This is an empirical question which needs to be examined. Lower sampling would reduce the reliability of the tests. Compounding this, the issue of pass marks must be addressed – compensation within the tests raises not only reliability questions but also washback effects into formative assessment. People who pass may still need to address key areas of learning in a key stage, if compensation and pass marks combine disadvantageously. The length of the tests and the need to cover the domain will tend to drive tests to a limited set of item types, raising validity issues. This in turn affects standards maintenance – if items are clustered around a text, if the text is changed (remembering test frequency is increased 100%) then all the items are no longer usable. This represents a dramatic escalation of burden in test development. Constructing and maintaining the bank of items will be very demanding.

xiv

If a high pass mark is set (and the facility of items tuned to this) there will be little evidence of what a child cannot do. Optimising the formative feedback element – including feedback for high attainers – in the face of demand for high domain coverage, reasonable facility, and accessibility (recognisable stimulus material etc) will be very demanding for test designers. Level-setting procedures are not clear. The regime requires a very fast turnaround in results – not least to set in place and deliver learning for a 're-take' in the next test session (as well as keeping up with the pace of progression through the National Curriculum content). This implies objective tests. However, some difficult factors then combine. The entry will be a volatile mix of takers and re-takers.

xv

While calibration data will exist for the items, random error will increase due to the volatility of entry, feeding into problems in the reliability of the item data in the bank. Put crudely, with no awarding processes (as present in existing national tests) there will be a loss of control over the overall test data – and thus reliability and standards over time will become increasingly problematic. As one possible solution, we recommend the development of parallel tests rather than successively different tests. Pre-tests and anchor tests become absolutely vital – and the purpose and function of these must be explained clearly to the public and the teaching profession. More information on this can be provided.

xvi

Having the same tests for different key stages (as stated by officials) is problematic. There is different content in different stages (see English in particular). QCA has undertaken previous work on 'does a level 4 mean something different in different key stages' – the conclusion was that it *did*.

xvii

The 10-hour training/learning sessions are likely to be narrowly devoted to the tests. This may communicate strong messages in the system regarding the importance of drilling and 'surface learning' – exactly the opposite of what the DfES is supporting in other policy documents. Although superficially in line with 'personalisation', It may instil dysfunctional learning styles.

xviii

We applaud the sensitivity of the analysis emerging from the DfES in respect of the different populations of learners who are failing to attain target levels. We also support the original Standards Unit's commitment to a culture of high expectations, combined with high support. However, this level of sensitivity of analysis is not reflected in the blanket expectation that every child should improve by two levels.

xix
We do not support 'payment by results' approaches – in almost any form these have successively been found wanting. Undue pressure is exerted on tests and test administration – maladministration issues escalate.

xx
In the face of the considerable challenge of developing a system which meets the demanding criteria which we associate with the operation of robust national assessment, we would welcome an opportunity to contribute to further discussions on the shape of enhanced national arrangements.

5. The way forward for National Assessment

58. What is needed is a new look at options – and both the technical and political space for manoeuvre. Cambridge Assessment has not only attempted to assemble the evidence but have produced a '3 option' paper which outlines possible approaches to confront the very real problems outlined above. We commend a thoroughgoing review of the evidence. Not a – 'single person review' like 'Dearing' or 'Tomlinson', but a more managed appraisal of options and a sober analysis of the benefits and deficits of alternatives. For this, we believe that a set of clear criteria should be used to drive the next phase of development:

- technically-robust arrangements should be developed
- the arrangements should be consistent with stated functions
- insights from trialling should fed into fully operational arrangements
- unintended consequences are identified and remedied
- full support from all levels of the system is secured in respect of revised arrangements
- a number of models should be explored at the same time, in carefully designed programmes – in other words there should be parallel rather than serial development, trialling and evaluation
- appropriate ethical safeguards and experimental protocols should be put in place during development and trialling

59. It is, of course, vital to consider not only the form of revised arrangements which better deliver the purposes of national assessment but also to consider the methods and time frame for development arrangements, as well as the means of securing genuine societal and system support.

60. The last two elements listed above are critical to this: currently, there are no plans for trialling more than one revised model for national testing. However, a cursory glance in the education research field shows that there is a range of contrasting approaches to delivering the key functions of national testing, many of which may well be presented to this Inquiry... It therefore would seem important to trial more than one model rather than 'put all eggs in one basket' or take forward only modifications of existing arrangements.

61. It is unclear whether adequate safeguards have been put in place to protect learners exposed to revised national assessment arrangements. Cambridge Assessment recommends – in line with the standards being developed by the Government's Social Research Unit – that new protocols should be developed, as a matter of urgency for the trialling of revised arrangements,.

National Assessment - Annexe 1

An overview of the evidence

1

Measurement error and the problems with overlaying levels onto marks

This does not refer to human error or mistakes in the administration of tests but to the issue of intrinsic measurement error. Contemporary standards in the US lead to the expectation that error estimates are printed alongside individual results: such as '…this person has 3592 (on a scale going to 5000 marks) and the error on this test occasion means that their true score lies between 3602 and 3582….'. Is this too difficult for people to handle (i.e. interpret)? In the current climate of increasing statistical literacy in schools, it should not be. Indeed, results could be presented in many innovative ways which better convey where the 'true score' of someone lies.

Error data are not provided for national tests in England, and both the Statistics Commission and commentators (e.g. Wiliam, Newton, Oates, Tymms) have raised questions as to why this international best practice is not adopted.

Of course, error can be reduced by changing the assessment processes – which most often results in a dramatic increase in costs. Note 'reduce' not 'remove' – the latter is unfeasible in mass systems. For example, double marking might be adopted and would increase the technical robustness of the assessments. However, this is impractical in respect of timeframes; it is already difficult to maintain existing numbers of markers, etc. Error can be reduced by increased expenditure but is escalating cost appropriate in the current public sector policy climate?

One key point to bear in mind is that one must avoid a situation where the error is significantly less than the performance gains which one is expecting from the system, and from schools – and indeed from teachers within the schools. Unfortunately, 1-2% improvement lies within the bounds of error – get the level thresholds wrong by two marks either way (and see the section on KS3 Science below) and the results of 16,000 pupils (i.e. just over 2%). could be moved.

Measurement error becomes highly significant when national curriculum levels (or any other grade scale) is overlaid onto the scores. If the error is as above, but a cut score for a crucial level is 48 (out of 120 total available marks) then getting 47 (error range 45-49) would not qualify that person for the higher level, even though the error means that their true score could easily be above the level threshold. In some cases the tests are not long enough to provide information to justify choosing cut-scores between adjacent marks even though the difference between adjacent marks can have a significant effect on the percentages of the cohort achieving particular levels. There are problems with misclassification of levels applied. Wiliam reports that 'it is likely that the proportion of students awarded a level higher or lower than they should be because of the unreliability of the tests is at least 30% at key stage 2 and may be as high as 40% at key stage 3'.

Criterion referencing fails to work well since question difficulty is not solely determined by curriculum content. It can also be affected by 'process difficulty' and/or 'question or stimulus difficulty', (Pollitt et al). It is also difficult to allocate curriculum to levels since questions testing the same content can cover a wide range of difficulty.

It is believed that error could be communicated meaningfully to schools, children, parents and the press, and would enhance both intelligence to ministers and the educational use of the data from national tests.

The current practice of overlaying levels onto the scores brings serious problems and it is clear that the use of levels should be reviewed. One key issue: consider the following:

Level 5 upper boundary ----------
Child C

Child B
Level 4 upper boundary ----------
Child A

Level 3 upper boundary ----------

Both Child B and Child C are level 5. But in fact Child A and B are closer in performance, despite A being level 4 and B being level 5. Further, if Child A progresses to the position of Child B over a period of learning, they have increased by one level. However, if Child B progresses to the same position as Child C, they have progressed further than Child A over the same time, but they do not move up a level. Introducing sub-levels has helped in some ways (4a, 4b etc) but the essential problem remains.

2
QCA practice in test development

Pretesting items is extremely helpful; it enables the performance characteristics of each item to be established (particularly how relatively hard or easy the item is). This is vital when going into the summer levels setting exercise – it is known what is being dealt with in setting the mark thresholds at each level. But subject officers and others involved in the management of the tests have had a history of continuing to change items after the second pretest, which compromises the data available to the level setting process, and thus impacts on maintaining standards over time. In addition, the 'pretest effect' also remains in evidence – learners are not necessarily as motivated when taking 'non-live' tests; they may not be adequately prepared for the specific content of the tests; etc. This places a limit on the pre-test as an infallible predictor of the performance of the live test.

3

Borderlining

The decision was taken early in national assessment to re-mark all candidates who fall near to a level threshold. QCA publish the mark range which qualifies children to a re-mark. However, the procedure has been applied only to those below the threshold and who might move up, and not to those just above, who might move down. This has had a very distorting effect on the distributions. Although done in the name of fairness, the practice is seriously flawed. For years, arguments around changing the procedure or removing borderlining completely foundered on the fact that this would effect a major (downward) shift in the numbers gaining each level, and therefore could not be sanctioned politically. A poorly-designed and distorting practice therefore continued. Current practice is unjustifiable and would not be sanctioned in other areas of public awarding (e.g. GCSE and A/AS).

It has now been agreed between QCA and the DfES that borderlining will be removed in 2008, when the marking contract passes from Pearson's to the American ETS organization. At this point, a recalibration of standards could be effected to mask the effect of correction and this standard could be carried forward, or a clear declaration could be made on how removal of borderlining affects the 'fairness' of the test and has resulted in a change in the numbers attaining a given level. First identified by Quinlan and Scharaskin in 1999, this issue has been a long-running systemic problem. Again, it is a change in practice (alongside changes in the accessibility of the tests, in inclusion of mental arithmetic etc) which compromises the ability of the tests to track change in attainment standards over time.

4

Fluctuations in Science at KS3

At levels 6 and above, standards of attainment have moved up and down in an implausible fashion:

2005	**37**	(% of children gaining levels 6 and 7)
2004	**35**	
2003	**40**	
2002	**34**	
2001	**33**	

The movement over the three year period 2002 to 2004 has involved a 6% increase followed by a 5% decrease – a movement of 11% over two years. This is implausible, and points to problems in the tests and level setting, and not to a real change in underlying standards or in the cohort taking the tests. Significantly, when interviewed on causes, officials and officers gave very different explanations for the effect – in other words, the true cause has not been established with precision.

5

The Massey Report and Tymms' analysis

The Massey report used highly robust method to triangulate national tests 1996-2001 and yields solid evidence that attainment standards have risen over that period, but not to the extent in all subjects and all key stages that has been argued by DfES and ministers. Tymms' less robust method and research synthesis suggests broadly the same. Massey made a series of recommendations, some of which have been adopted by QCA, such as equating a number of years' tests and not just the preceding year. However, the absence of a consistent triangulation method and the failure to adopt the Massey recommendation that standards should be held for five years and then publicly recalibrated has not been adopted.

6

Ofsted's over-dependence on national test outcomes

The new Ofsted inspection regime is far more dependent on the use of national assessment data than previously. This delivers putative economies since Ofsted feels it can better identify problematic and successful schools, and can use the data to target areas of schools – e.g. weak maths departments, or poor science etc. The revised regime is broadly welcomed by schools, and has a sound emphasis on each school delivering on its stated policies. But the regime fails to acknowledge the weaknesses of the data which lie at the heart of the pre-inspection reports, and which guides Ofsted on the performance of schools. The greatly increased structural dependence on data which is far less accurate than is implied is problematic. The new regime delivers some valuable functions – but the misapprehension of the real technical rigour of the assessment data is a very serious flaw in arrangements.

7

Assessment overload accusations whilst using many other non-statutory tests

This is an interesting phenomenon – the optional tests are liked, the statutory tests are frequently disliked (QCA). KS2 score are mistrusted (ATL). The use of 'commercial' CAT tests and CEM's tests (MIDYIS etc) is widespread. CAT scores are trusted by teachers because the results are more stable over time in comparison with national curriculum tests; this reflects the different purpose of the respective instruments. Children say 'I did SATs today' when they do a statutory key stage test. They also frequently say that when they have taken a CAT test. There is widespread misunderstanding of the purpose of the range of tests which are used. QCA was lobbied over a five-year period to produce guidance on the function of different tests – not least to clarify the exact purpose of national testing. However, no such guidance has been produced. As a result of this, the arguments regarding 'over-testing' are extremely confused, and adversely muddy the waters in respect of policy.

8

Is the timing right?

Changing the timing of the tests would require a change in primary legislation. However, it is an enhancement of testing which should be considered very seriously. In the final report of the Assessment Review Group in Wales, Daugherty (2004) recommends that 'serious consideration should be given to changing the timing of Key Stage 3 statutory assessment so that it is completed no later than the middle of the second term of Year 9'. The Group believed the current timing to be unhelpful in relation to a process that could, in principle, inform,' and that, 'one source of information that would be of use potentially to pupils and their parents is not available until after the choice of pathway for Year 10 and beyond has been made'. There are also implications for the potential use of Key Stage 1 and 2 data for transition between phases. 'School ownership' – taking the outcomes very seriously in managing learning – would be likely to increase in this re-scheduling of the tests.

9

The reliability of teacher assessment

Particularly in the high stakes context of performance tables, we feel that relying on teacher assessment, as currently operated, is not a robust option. Work in 2000 by QCA Research Team showed a completely unstable relationship between TA and test scores over time at school level. This is compelling evidence against an over-dependence on teacher assessment. There are means of delivering moderated teacher assessment for reporting to parents, and bolstering accountability not by testing but by regional inspection based on high expectations and school improvement models (see recommendations below). National standards in underlying attainment could be delivered through a light sampling model (with matrix sampling to cover all key content of the national curriculum). This would enable a valid answer to the ministerial question '….nationally, what's happening to standards in English?'.

10

Teaching to the test

The recent lobbying by Baroness Professor Susan Greenfield and eminent colleagues is merely the most recent critique of the problems of teaching to the test. The 'Texas Test Effect' (Wiliam, Oates) is well known but poorly presented to Government. Bill Boyle (CFAS) is the latest empirical study of the adverse effects of teaching to the test and its almost universal domination of educational purposes in the English school system. It is a very serious issue, and it may be one significant factor (not the sole one) lying behind the 'plateau effect' associated with the majority of innovations such as the Primary Literacy and Numeracy Strategies. In other words – a succession of well-intended and seemingly robust initiatives repeatedly run out of steam.

National Assessment - Annexe 2
The Assessment of Performance Unit – should it be re-instated?

The origins and demise of the APU

1. The inability of current arrangements to provide a robust flow of policy intelligence on trends in pupil attainment has emerged as a serious problem. The causes are multifaceted, and include :

- instability in standards within the testing system (Massey, Oates, Stats Commission)
- acute classification error affecting assignment of pupils to levels (Wiliam, Tymms)
- teaching to the test/'Texas Test Effect' (Wiliam)

2. Growing awareness of this issue has prompted increasing calls for '...a return to the APU…' (the Assessment of Performance Unit) – a separate, 'low stakes', light-sampling survey for the purpose of reliable detection of patterns of pupil attainment, and of trends in attainment over time. But there are dangers in an unreflective attempt to re-instate arrangements which actually fell short of their aims. The APU processes were innovative and progressive. They mirrored the fore-running US NAEP (National Assessment of Educational Progress) and pre-dated the arrangements now in place in New Zealand and Scotland. Running surveys from 1978 to 1988, politicians and civil servants saw it being redundant in the face of the data on each and every child of age 7,11 and 14 which would be yielded from National Curriculum assessment processes. The APU was hardly problem-free. Significant issues emerged in respect of:

- under-developed sampling frames
- tensions between subject-level and component-level analysis and reporting
- differing measurement models at different times in different subjects
- lack of stability in item forms
- escalating sampling burden
- difficulty in developing items for the highest attaining pupils
- the nature of reporting arrangements
- replacement strategy in respect of dated items
- ambiguity in purpose re 'process skills' as a principal focus versus curriculum content
- research/monitoring tensions
- compressed development schedules resulting from pressure from Government

3. There was acute pressure on the APU to deliver . Rather than recognize that the function of the APU was essential for high-level policy processes and would need to persist (NEAP has been in place in the US since 1969), the intense pressure led to poor refinement of the technical processes which underpinned the operation of the APU, and high turnover in the staff of the different subject teams (Gipps and Goldstein 1983). Crucially, the compressed piloting phases had a particularly adverse impact; there was no means of undertaking secure evaluation of initial survey work and feeding in 'lessons learned':

> '…the mathematics Group, in particular, felt that they were continually being rushed: their requests for a delay in the monitoring programme were rejected; their desire for three pilot surveys was realized as only one; they experienced a high turnover of staff and a resulting shortage of personnel. The constant rush meant that there was no time for identifying and remedying problems identified in the first year of testing.

4. In fact, all three teams suffered from a rapid turnover of staff, put down to the constant pressure of work combined with a lack of opportunity was no time to 'side track' into interesting research issues…' (Newton P 2005 p14)

5. This is not a trivial failure of a minor survey instrument. The APU was founded in 1974 after publication of a DES White Paper (Educational Disadvantage and the Needs of Immigrants). It was the result of a protracted strategic development process, which led from the DES-funded Working Group on the Measurement of Educational Attainment (commissioned in 1970) and the NFER's DES-funded development work on Tests of Attainment in Mathematics in Schools. If it had successfully attained its objectives, it would have relieved National Curriculum testing of the burden of attempting to measure standards over time – a purpose which has produced some of the most intense tensions amongst the set of functions now attached to national testing. Stability in the instruments is one of the strongest recommendations emerging from projects designed to monitor standards over time. In sharp tension with this, QCA and the State has – in line with commitments to high quality educational provision; the standards agenda; and responses from review and evaluation processes - sought to optimize the National Curriculum by successive revision of content; increasing the 'accessibility of tests'; and ensuring tight linkage of the tests to specific curriculum content. These are laudable aims – and the emphasis on the diagnostic function of the data from tests has been increasing in recent innovations in testing arrangements. But pursuit of these aims has led to repeated revision rather than stability in the tests.

6. The Massey Report suggested that if maintenance of standards over time remained a key operational aim, then stability in the test content was imperative. In the face of these tensions, retaining an APU-style light sampling survey method would enable de-coupling of national assessment from a requirement to deliver robust information on national educational standards, and enable testing to reflect curriculum change with precision, to optimize the learning-focussed functions of testing, and enable constant innovation in the form of tests (e.g. to optimize accessibility).

7. Thus, the deficits and closure of the APU were, and remain, very serious issues in the operation and structure of national assessment arrangements. Temporal discontinuity played a key role in the methodological and technical problems experienced by the APU developers. As outlined above, rushing the development phases had a variety of effects, but the most serious of these was the failure to establish with precision a clear set of baseline data, accompanied by stable tests with known performance data; '…an effective national monitoring system cannot be brought 'on stream' in just a couple of years…' (Newton P, 2005).

8. Our conclusion is not 'bring back the APU', but develop a new light sampling, matrix-based model using the knowledge from systems used in other nations and insights from the problems of the APU. Models 1 and 2 which we outline as alternatives in the main body of this evidence rely on the development of new versions of the APU rather than simple re-instatement.

Section 2
5. Higher education admissions tests

Determining role and function

1. Since the publication, in September 2001, of the Schwartz report (Fair Admissions to higher education: recommendations for good practice), the issue of the role and function of admissions tests has been a controversial area. Cambridge Assessment has been cautious in its approach to this field. We have based our development programme on carefully-considered criteria. We believe that dedicated admissions tests should:

- produce information which does not duplicate information from other assessments and qualifications
- make a unique and useful contribution to the information available to those making admissions decisions
- predict students' capacity to do well in, and benefit from higher education

2. Since the Cambridge Assessment Group includes the OCR awarding body, we are also heavily involved in refining A levels in the light of the 'stretch and challenge' agenda – working to include A* grades in A levels, inclusion of more challenging questions, and furnishing unit and UMS scores (Uniform Mark Scheme scores –a mechanism for equating scores from different modules/units of achievement) as a means of helping universities in the admissions process.

3. We recognize that HE institutions have clear interests in identifying, with reasonable precision and economy, those students who are most likely to benefit from specific courses, are likely to do well, and who are unlikely to drop out of the course. We also recognize that there is a strong impetus behind the 'widening participation' agenda.

4. Even with the proposed refinements in A level and the move to post- qualification applications (PQA), our extensive development work and consultation with HE institutions has identified a continuing need for dedicated assessment instruments which facilitate effective discrimination between high attaining students and are also able to identify those students who possess potential, but who have attained lower qualification grades for a number of reasons.

5. We are very concerned not to contribute to any unnecessary proliferation of tests and so have been careful only to develop tests where they make a unique and robust contribution to the admissions process, enhance the admissions process, and do not replicate information from any other source. To these ends, we have developed the BMAT for medical and veterinary admissions. We have developed the TSA (Thinking Skills Assessment), which is being used for admissions to some subjects in Cambridge and Oxford and is being considered by a range of other institutions. The TSA items (questions) also form part of the uniTEST which were developed in conjunction with ACER (Australian Council for Educational Research). UniTEST is being trialled with a range of institutions, both 'selecting' universities and 'recruiting' universities.

6. This test is designed to help specifically with the widening participation agenda. Preliminary data suggests that this test is useful in helping identify students who are capable of enrolling on courses at more prestigious universities than the ones for which they have applied as well as those who should consider HE despite low qualification results.

7. The TSA should be seen more as a test resource rather than a specific test: TSA items are held in an 'item bank', and this is used to generate tests for different institutions. Although TSA items were originally developed for admissions processes in Cambridge where discrimination between very high attaining students is problematic and A level outcomes inadequate as a basis for admissions decisions, Cambridge Assessment research team is developing an 'adaptive TSA'. This utilizes the latest measurement models and test management algorithms to create tests which are useful with a very broad range of abilities.

8. The validation data for the TSA items is building into a large body of evidence and the tests are yielding correlations which suggest that they are both valid and useful in admissions - and do not replicate information from GCSE and AS/A2 qualifications. In other words, they are a useful addition to information from these qualifications and allow more discriminating decisions to be made than when using information from those qualifications alone. In addition, they yield information which is more reliable than the decisions which are made through interviews and will provide a stable measure over the period that there are major changes to AS and A2 qualifications.

The American SAT

9. Cambridge Assessment supports the principles which are being promoted by the Sutton Trust and the Government in respect of widening participation. However, we have undertaken evaluation work which suggests that the promotion of the American SAT test as a general admissions test for the UK is ill-founded. The five-year SAT trial in the UK is part-funded by Government (£800,000), the College Board (the test developers) contributing £400,000 and with the Sutton Trust and NFER each contributing £200,000.

10. The literature on the SAT trial in the UK states that the SAT1 is an 'aptitude' test. It also makes two strong claims that are contested:

> "…Other selection tests are used by universities in the United Kingdom, but none of these is as well constructed or established as the SAT©.
>
> In summary, a review of existing research indicates that the SAT© (or similar reasoning-type aptitude test) adds some predictive power to school / examination grades, but the extent of its value in this respect varies across studies. In the UK, it has been shown that the SAT© is an appropriate test to use and that it is modestly associated with A-level grades whilst assessing a different construct. No recent study of the predictive power of SAT© results for university outcomes has been undertaken in the UK, and this proposal aims to provide such information…"
>
> Source: (http://www.nfer.org.uk/research-areas/pims-data/outlines/update-for-students-taking-part-in-this-research/a-validity-study-background.cfm)

11. The claim that 'none of these is as well constructed or established as the SAT© ' fails to recognise that Cambridge Assessment has assembled comprehensive data on specific tests amongst its suite of admissions tests and ensures that validity is at the heart of the instruments. These are certainly not as old as the SAT but it is entirely inappropriate to conflate quality of construction and duration of use.

12. More importantly, the analysis below suggests that the claim that the SAT1 is a curriculum-independent 'aptitude' test is deeply flawed. This is not the first time that this claim has been contested (Jencks, C. and Crouse, J; Wolf A and Bakker S), but it is that first time that such a critique has been based on an empirical study of content.

13. It is important to note that the SAT is under serious criticism in the US (Cruz R; New York Times) and also, despite many UK-commentators' assumptions, the SAT1 is not the sole, or pre-eminent, test used as part of US HE admissions (Wolf A and Bakker S). The SAT2 is increasingly used – this is an avowedly curriculum-based test. Similarly, there has been a substantial increase in the use of the Advanced Placement Scheme – subject-based courses and tests which improve students' grounding in specific subjects, and are broadly equivalent to English Advanced Level subject-specific qualifications.

14. It is also important to note that (i) the US does not have standard national examinations – in the absence of national GCSE-type qualifications, a curriculum-linked test such as the SAT1 is a sensible instrument to have in the US, to guarantee that learners have certain fundamental skills and knowledge – but GCSE fulfils this purpose in England; (ii) the USA has a four-year degree structure, with a 'levelling' general curriculum for the first year; and (iii) the SAT1 scores are used alongside college grades, personal references, SAT2 scores and Advanced Placement outcomes.

> "…One of the misunderstood features of college selection in America is that SATs are only one component, with high school grades and other 'portfolio' evidence playing a major role. The evidence is that high school grades are a slightly better predictor of college achievement than SAT scores, particularly for females and minority students. Combining both provides the best, though still limited, prediction of success…" (Stobart G)

Curriculum mapping – does the SAT mirror current arrangements?

15. In the light of research comment on the SAT and emerging serious criticisms of the instrument in its home context, Cambridge Assessment commissioned a curriculum mapping of the SAT in 2006 comparing it with content in the National Curriculum (and, by extension, GCSE) and the uniTEST.

16. It is surprising that such a curriculum content mapping has not been completed previously. Prior studies (McDonald et al) have focused on comparison of outcomes data from the SAT and qualifications (e.g. A level) in order to infer whether the SAT is measuring something similar or different to those qualifications. But the failure to undertake a comparison of the SAT with the content of the English National Curriculum is a serious oversight. The comparison is highly revealing.

17. The study consisted of a comparison of published SAT assessment criteria, items included in SAT1 sample papers, the National Curriculum programmes of study, and items within the uniTEST. The SAT assessment criteria and National Curriculum programmes of study were checked for analogous content. The National Curriculum reference of any seemingly relevant content was then noted and checked against appropriate SAT1 specimen items. The full analysis was then verified by researchers outside the admissions team, who were fully acquainted with the content of the National Curriculum and GCSEs designed to assess National Curriculum content. The researchers endorsed the analysis completed by the admissions test developers.

The outcomes of the curriculum mapping study

18. The full results are shown in Higher education admissions tests annexe #1. Column 1 shows the sections and item content of the SAT1. Column 2 gives the reference number of the related National Curriculum content. For example, MA3 2i refers to the statement:

> Mathematics key stage 4 foundation
> **Ma3 Shape, space and measures**
>
> Geometrical reasoning 2
> **Properties of circles**
>
> recall the definition of a circle and the meaning of related terms, including centre, radius, chord, diameter, circumference, tangent, arc, sector, and segment; understand that inscribed regular polygons can be constructed by equal division of a circle

19. Column 3 in Annex #1 shows the relation between the content of the SAT1, the relevant components of the National Curriculum and the Cambridge/ACER uniTEST admissions test.

20. The analysis indicates that

- the SAT1 content is largely pitched at GCSE-level curriculum content in English and Maths, and replicates GCSE assessment of that content.
- the item types and item content in the SAT1 are very similar to that of GCSEs.

It is therefore not clear exactly what the SAT1 is contributing to assessment information already generated by the national examinations system in England.

21. Previous appraisals of the SAT1 have been based on correlations between GCSE, A level and SAT1 outcomes. This has shown less than perfect correlation, which has been interpreted as indicating that the SAT1 assesses something different to GCSE and A level. But GCSE and A level are based on compensation – particularly at lower grades, the same grade can be obtained by two candidates with different profiles of performance. The inferences from the data were previously made in the absence of a curriculum mapping. The mapping suggests that discrepancies between SAT1 and GCSE/A level outcomes may be the result of the candidates not succeeding at certain areas in these exams, nonetheless gaining a reasonable grade – but this being re-assessed by SAT1 and thus their performance found wanting.

22. The existence of such comprehensive overlap suggests that the SAT either presents an unnecessary replication of GCSE assessment or an indication of the problems of compensation in the established grading arrangements for GCSE.

23. Identical analysis of uniTEST, currently being piloted and developed by Cambridge Assessment and ACER, suggests that uniTEST does not replicate GCSE assessment to the same extent as the SAT1 but focuses on the underlying thinking skills rather than on formal curriculum content. There is some overlap in the areas of verbal reasoning, problem solving, and quantitative and formal reasoning. There are, however, substantial areas of content which are not covered in the National Curriculum statements of attainment nor in the SAT1. These are in critical reasoning and socio-cultural understanding. This suggests that uniTEST is not replicating GCSE and does offer unique measurement. Preliminary data from the pilot suggest that uniTEST is detecting learners who might aspire to universities of higher ranking that the ones to which they have actually applied.