

International Approaches to Science Assessment

Paul Black, King's College London

Dylan Wiliam, ETS

Paper commissioned by the Committee on Test Design for K-12 Science Achievement
Center for Education
National Research Council

Copyright © 2004 National Academy of Sciences. All rights reserved. No part of these pages, either text or image may be used for any purpose other than personal use. Reproduction, modification, storage in a retrieval system or retransmission, in any form or by any means, electronic, mechanical or otherwise, for reasons other than personal use, is strictly prohibited without prior written permission.

Opinions and statements included in the draft papers are solely those of the individual author(s), and are not necessarily adopted or endorsed or verified as accurate by the Committee on Test Design for K-12 Science Achievement or the National Academy of Sciences, including the National Academy of Engineering, Institute of Medicine, or National Research Council.

Introduction

Over the last ten years, a number of analyses of the assessments used in different countries have appeared. Eckstein and Noah (1993) reviewed the experiences of students in the final year of upper secondary education in China, England and Wales, France, Germany, Japan, Sweden, the Soviet Union, and the United States, looking at the details of the assessments used, and the wider policy issues, such as success rates and the degree of centralization. Britton and Raizen (1996) looked in greater detail at the assessments used in science and mathematics, and in particular at the coverage of different topics. These, and other analyses emerging from international comparisons such as TIMSS and PISA, have focused on what might be termed 'cross-sectional' comparisons; examining the differences between the assessments in different national systems for students of the same age. Other studies have focused on the study of systems of assessments within a single country, looking at the articulation between the various assessments taken by students at the same ages (another form of cross-sectional analysis within a single system), and by the sequence of assessments taken by an individual over the course of their schooling (a longitudinal focus within a single system).

The purpose of this paper is to try and combine both approaches, in the form of a comparison of assessment *systems*. Our aim is to look at the differences in assessment systems in different countries in order to try to identify the critical design issues. In this context, design issues are not just the features that are designed into the system by its architects. Design issues also include those features of the system that emerge unplanned, but might have panned out differently if different decisions had been taken in the development of the system (good examples of this latter category are the issue of the rate of increase of spread of attainment in the cohort over time, or sex differences in performance).

In writing this paper, we have assembled a series of brief outlines of the assessment systems in **seven different countries**: Australia (Queensland), France, Germany, Japan, New Zealand, Sweden and the United Kingdom (England). These systems were chosen because individually, they each differ in important ways from practice in the United States, and collectively they illustrate a range of assessment practices in use for large-scale assessment. From these country descriptions, we have drawn out eight major areas that we believe need to be considered in designing an assessment system for science for grades K-12 in the United States. These issues are:

1. The purposes of assessment

In this section, we discuss the different functions, including supporting learning, attesting to the accomplishments, capabilities and potential of individual learners, and holding institutions to account.

2. The structure of the assessment system

This section outlines some of the issues involved in deciding upon an ‘architecture’ for an assessment system, such as the articulation between the assessments taken by students at different ages, the way that achievement is reported, and how the way that the subject is defined interacts with these issues.

3. The locus of assessment

Here, we discuss the issues of who creates the assessments, by whom, when and where they are administered, and who scores them.

4. Extensiveness

In this section, the focus is on who is assessed and on what basis. In particular, this section discusses the choice between sampling and census approaches, and the extent to which special populations are excluded from the assessment, or receive modified forms of the assessment.

5. Assessment format

This section outlines very briefly the extensive literature on assessment formats, such as the multiple-choice versus constructed-response items, the role of portfolios, and other kinds of assessments.

6. Scoring models

The way that scores on individual items are combined, aggregated, reconciled and reported are often taken for granted, and yet the way that such issues are settled can have far-reaching consequences for the performance and the effects of an assessment system, such as the meanings that can be attached to grades or scores. In this section, these issues are outlined briefly, and some alternative models for scoring are discussed.

7. Quality issues

In this section we discuss the quality of assessments under the general heading of validity, within which we discuss important threats to quality of assessment such as bias, unfairness, and unreliability.

8. The role of teachers

A common theme arising in sections 1 to 7 is the role of teachers in assessment, and in this section, this issue is discussed in greater detail.

9. Contextual issues

This final section analyzes the technical issues discussed in terms of their relationship to broader issues such as consistency with findings from other fields within the applied social sciences such as psychology, sociology, and policy.

For the sake of readability, we present the eight thematic sections first, followed by general references, and then we present the outlines of the seven assessment systems. Further readings on each of the systems are provided immediately after the description of the system.

1. The purposes of assessment and testing

Three purposes

The main types of purpose are concerned respectively with the *support of learning*, with *certification*, i.e. with reporting the achievements of individuals, and with satisfying demands for public *accountability*. A distinction has to be made at the outset between these purposes and the instruments and procedures that might be used to serve them. For example, the same test questions may be used for quite different purposes, and, conversely, a single purpose might be served by combining the results obtained from a range of different types of assessment.

For learning

When assessment is *formative*, it shapes learning. For example, information evoked in day-to day classroom work can be used, typically in the immediate to short-term, to modify the learning and teaching work. The research evidence shows that this assessment can be a powerful support for learning (Black and Wiliam 1998) and ways in which it can be implemented in practice have been developed (e.g. Black et al. 2002, 2003). Such assessment makes teaching more closely interactive and adaptive so that its development implies far more than addition of a few new procedures to an existing, non-interactive, classroom practice. The experience of the King's team fully bears out the analysis of Linn (1989) who pointed out that:

the design of tests useful for the instructional decisions made in the classroom requires an integration of testing and instruction. It also requires a clear conception of the curriculum, the goals, and the process of instruction. And it requires a theory of instruction and learning and a much better understanding of the cognitive processes of learners.

The practice of formative assessment can be productive and rewarding for both students and teachers, but it has to be informed by a model that is quite detailed, in that it has to provide some guidance about the ways in which a pupil might progress in learning, linked to a clear conception of the curriculum and its learning goals.

For certification

The need for *certification* of students can arise both within schools, between schools and when students leave schooling altogether. For transfer between different stages of schooling and between different teachers any certification assessment has to be an effective communication, in that the information has to be formulated with a structure and a language that reflects a shared understanding between those who are communicating. Different teachers have to be working to common standards and have to understand one another's procedures for determining standards of grading. The purpose implies that teachers teaching the same school subject have to be constrained by a shared system. The purpose might be better served by a profile detailing a pattern of strengths and weaknesses rather than by a single aggregate grade.

For transfer between and out of schools, similar constraints apply. However, communication of criteria and standards is less informal, so clear and agreed documentation is essential, and the information will be worthless unless different schools work to common schemes and produce assessments on a shared basis, or within a common external scheme. Such considerations bear most heavily on the large discontinuities involved when pupils move either out of school into employment, or to further and more advanced study. Here, in addition to the pressure to ensure overall comparability, there is pressure to ensure the validity of any system.

In general, the certification purpose of assessment raises two further issues. One is that the decisions to be taken about who is to control the process and who is to carry it out. Here are included choices about the local or national control, about whether this control is general and flexible or close and detailed, and about the assignment of responsibility, whether to teachers and schools, or to external agencies, or to a combination of these. These issues are further discussed below in Section 3.

A second issue is that decisions have to take account of the costs implied. The cost of public examinations is a significant item in the budgets of many different stakeholders; for schools, high-stakes summative assessments which have to work to shared external standards bear a high cost in terms of the time of classroom teachers and of school managements. On the other hand, complete reliance on forms of assessment that can be delivered at low cost will almost certainly result in only a small proportion of what is valued being assessed.

For accountability

This purpose can raise the most severe problems. Whilst schools and colleges clearly have a responsibility to the public who fund them, the data on performance outcomes derived from assessment and testing cannot on their own provide valid guidance. A wide range of data about the backgrounds of students who enter a school or college, and about resources, is needed if the achievements of the students in any one school are to be interpreted in ways that are useful for policy purposes and to help schools themselves to improve. In some countries, systems of inspection by trained teams who visit a school to audit its work also serve an important role, but such a system can be either primarily supportive, or judgmental and coercive.

There is also a need to examine the assumption that the accountability function is best met by collection of the certification data from all students, thus conflating the certification and accountability functions. To produce an overall picture of a nation's performance, it is not necessary to test every student within a given age group. Indeed, using matrix light sampling, it is possible for such surveys to explore a wider range of attainments in greater detail than would be possible with a test which would have to be the same for every pupil. That such surveys can produce results useful to teachers in their work has been shown by surveys in the UK in the 1980s (Johnson, 1989), and in the current New Zealand system. However, if individual schools are to be judged, then requirements of sample size might mean, in the case of small schools, that most students have to be tested.

Synergies and tensions

The choices which will distinguish or adapt assessments to their purposes may affect the particular set of test items or procedures chosen, the way in which these are administered and marked and the ways in which the outcomes may be analyzed, combined and, particularly, interpreted according to the purpose given priority.

The three elements to be addressed here are assessment methods, assessment agencies and the purposes. One testing method carried out by a single agency might serve more than one purpose and would thereby be economical. If on the other hand, the methods to be used for different purposes have to be completely different, or the interests of different agencies are in tension or even opposed, then separation is required.

There are many possible cases of overlap, tension or synergy, between the different purposes. The problems of combining *certification* with *accountability* have already been mentioned. Another notable case is the possible linking between the *learning* and the *certification* purposes, which implies a linkage between formative and summative practices. The possibility here is that assessment by teachers might serve both the formative and summative purposes for their students and so contribute to summative assessments for certification and accountability. In some countries this is done by having the school's summative result published alongside the result of external testing, in others by combining teachers' assessments with external measures in arriving at the final

summative result (UK, Sweden). Others again (Queensland) rely entirely on a school's assessments and thereby remove the need for operation of separate agencies and procedures to serve the certification purpose, and perhaps also the need for separate performance data for accountability purposes. Some have emphasized the differences between formative and summative purposes, and have argued that the assessment instruments and procedures needed for the one are so different from those for the other that neither can flourish without clear separation. On the other side it can be argued that the two functions are two ends of the same spectrum and that there is no sharp difference, and that if the two functions are separated, then teachers' assessment work will be devalued. Recent experience in the UK has exposed how teachers who have developed good formative practices find that extension of such to improvement of the whole range of their assessment work is limited by the pressures of the national high-stakes tests (Black and Wiliam, 2004).

The formative-summative spectrum

At one end of the spectrum is the classroom assessment where the formative action is immediate. The time scale is short, so feedback is targeted and instant. The domain is small, so that a single item encompasses the entire domain, and therefore domain sampling is not an issue. With a little pedagogic skill, disclosure (the extent to which an item or assessment discloses the student's knowledge or skill) is also not a problem. Opportunity to Learn (OTL) is guaranteed and so reliability is a minor problem. Many criteria for validity are covered, but precisely because of the short time scale and the limited context, predictive validity is uncertain. The outcome would not be a good basis for any decisions other than those calling for immediate feedback.

Some way along the spectrum might lie the short class quiz. Feedback could still be rapid. The domain is larger, and cannot be assessed in its entirety, so sampling is an issue, and the shorter the test, the more strongly the sampling issues reduce the reliability. Disclosure is also more of a problem, although feedback work can clear up its problems. There is now also a memory/retention issue, but this could help with predictive validity. Feedback might bear on particular points (as for the purely short-term formative) or on revision/retention practices, or on synthesis across the domain. OTL may still be guaranteed, but if synthesis is part of the test aims, opportunities to learn and practice appropriate synthesis would be a pre-requisite (expecting students to establish for themselves the inter-connection of different ideas usually turns out to be hopelessly optimistic). The result is now a firmer basis for decision about the future, but only if the decision implies inferences that can fairly be based on the test's demands.

The argument can be developed by travelling further along the spectrum, through end-of-year assessments, and then end-of-course assessments, until one reaches a *reductio ad absurdum* (otherwise known as the Last Judgment?). To ask a Ph.D. candidate to sit B.S. finals again, as well as to undergo a thesis oral, would be ridiculous. But it does raise the question of the amount of retention, and the range of domain sampling beyond which the summative enterprise ceases to make sense.

The answer lies in the issue of validity. One validity issue is the extent to which the assessment exercises faithfully relate to the inferences to be made. But of relevance here are the other issues of timescale and retention, as made clear by the arguments for combining results of a sequence of assessments occurring at regular intervals throughout a course rather than relying on a single 'big-bang' terminal test. There is no universal argument here: "they could do it once so no doubt they could do it again if needed" applies to some future uses of what has been learned, but "you have to be able to remember/understand/act on the spot without time to look it up" applies to others. Most inferences about school test results are surely in the former category; for those clearly in the latter, constant exercise of the skills and judgments together with regular review of the learning is often required (e.g. refresher exercises for airline pilots).

Whilst reliability is an ever-present constraint which changes in its demands, what is more critical is the shift in validity as one moves across this spectrum, a shift from faithfulness to the immediate aims underlying the learning tasks, to extrapolation to inferences implied in the decisions to be made on the basis of the assessment result. Both aspects are involved everywhere, but their significance and relative importance changes across the spectrum. Both aspects can involve feedback to the learner. It would be sensible for a teacher, knowing the detail of a student's performance on (say) an advanced placement test, to give feedback and advice about how to repair omissions and proceed more effectively in future study/employment; it cannot happen if the teacher has little or no part in the assessment of the advanced placement test, so an opportunity is lost.

The separation of summative from formative has distorted the proper consideration of summative. One can highlight the difference as one between improving past/present learning and informing future learning, but these are merely different aspects of the same landscape, with the differences exacerbated by particular assessment policies. To regard summative as completely separate leads to various pathologies and inefficiencies.

Where summative becomes the hand-maiden of accountability matters can get even worse, for accountability is always in danger of ignoring its effects on learning and thereby of undermining the very aim, of improving schooling, that it claims to serve. Accountability can only avoid shooting itself in the foot if, in the priorities of assessment design, it comes *after* learning.

Comments on some differences between countries

One of the most remarkable things about looking at the countries under study here is the almost total absence of pattern. In France, Germany, Japan and the United Kingdom, student access to higher education is primarily on the basis of one set of examinations. In France and Germany, the results of these assessments are not used as the primary indices of the quality of the schools, whereas in Japan and the UK they are. In Australia, New Zealand and Sweden, the judgments made by teachers, sometimes built up over the years, are the major input into university selection, and external assessments serve primarily to bring the standards of teachers into line across subjects (Queensland) and across schools

(Sweden). However, in Sweden, students also have the option of making a case for admission to university on the basis of scores on an aptitude test modeled on the College Board's SAT.

The system for university entrance in the USA is similar to that in Sweden, in that both teachers' judgments (in the form of grade-point average) and external test scores (from the SAT, ACT and advanced placement tests) are used. However, in its ranking of schools for the purposes of public accountability, it resembles more the United Kingdom or Japan¹. Perhaps the only strong finding from the countries studied here is that while many systems rely on teacher judgment for assessments that are high-stakes for students, there are, perhaps not surprisingly, no systems that rely on teacher judgment for assessments that are high-stakes for teachers.

New Zealand was remarkable in trying to use teachers' assessments and the unit standards to escape radically from the 'big-bang' model in which the decisions about university entrance are made on the basis of performance in one set of tests or examinations. Feasibility, plus conservatism, forced a compromise, but the achievement standards model still bears some similarities to a 'graduated assessment' model. Accountability is made comprehensive (i.e. all aspects of the school's work), and monitoring made to serve learning validity as well as to audit.

France is likewise strong on making the support of learning, through support of the teaching profession, a clear priority, whilst using sampling audits and broader inspections of schools for accountability. However, it retains its 'gold-standard', the Baccalaureat (or sometimes just 'Bac') as its 'big-bang'. Somewhat surprisingly, a few 'big' questions, rather homogeneous in style, are used for the summative purpose— issues of domain sampling, both in topic coverage and over questioning styles, seem to cause little or no concern; however, the involvement of teachers and the possibility of oral exams may ameliorate some of the shortcomings.

Germany is similar to France on the issue of terminal examinations, with some regions delegating more to a students' teachers. What is unique is that on their journey through education students' meet high-stakes decisions every year; these are made mainly by their school and appear to be high-stakes for the student, but not for the school. However, their recurring presence must mean that the summative function predominates (and may account for the dominance of formal teaching). One could well wonder why a strong searchlight is not directed on the validity and reliability of the evidence on which these decisions are based, particularly for the decision as to future secondary school taken at the end of the primary phase.

¹ This is particularly ironic since the Japanese and UK assessments measure subject knowledge acquired in school, which is therefore at least somewhat within the school's control, while the whole reason for the original development of the SAT was to minimize the contribution of preparation on students' scores (Lemann, 1999).

The tradition in the USA is quite different in the strong influence of the use of multiple choice and well-developed psychometrics, with the need to demonstrate high levels of reliability tending to take precedence over validity. The lack of an agreed or mandated national curriculum means that commercially available tests tend to focus on the ‘lowest common denominator’ of state and district content standards, thus raising substantial concerns regarding the alignment of standards, instruction, and assessment (in this context, it should be noted that ‘alignment’ is not an issue in countries like France or the UK that use final examinations, since they tend to shape the curriculum). While teachers’ own assessments of their students’ achievement do figure in admission to university via the grade-point average, the tendency has been to ‘work around’ teachers, and use externally-set tests as more ‘objective’ measures of student potential, and so there has been little development of teachers’ ability to assess their students’ achievement against agreed standards. Teachers’ assessment skills therefore appear to be rather rudimentary and would require considerable development before one could have faith in teachers’ judgments as sources of adequately reliable and valid scores.

The UK is unique in having a foot in all of the camps, having more lengthy and expensive tests – particularly for university entrance – because of trying to have the advantages of constructed-response items, with a dash of school-based assessment thrown in, whilst at the same time using the outcomes for the purpose of accountability. Note however that in England (a) there is now a trial, with a substantial sample of primary schools, to replace the tests of 7-year-olds by classroom assessment carried out by teachers, (b) a similar trial for doing the same with the tests of 14-year-olds is being explored; and (c) a major review of assessment policy at 18+ under the chairmanship of Mike Tomlinson is considering replacing the current academic and vocational examinations by a system which looks at this early stage remarkably similar to that operating in New Zealand. In Scotland the testing system in England has never been adopted and current initiatives focus mainly on developing formative assessments; using new-found independence, Wales is abandoning the system used in England (that it followed until recently), whilst Northern Ireland also plans to go its own way.

2. The structure of assessment systems

This section deals with some of the structural and technical issues relating to the design of assessment systems that are highlighted by the examination of assessment systems in other countries.

Vertical equating

One of the most important, and most neglected, issues in the design of assessment systems is the articulation between assessments administered to students of different ages. Driven by the need to provide high-quality assessments for students at particular points in their education careers, most of the research undertaken by psychometricians in the last 50 years has focused on the relationships between assessments taken by students of approximately the same age. For example, between April 1942 and May 1969, a total

of 82 different versions of the Scholastic Aptitude Test (produced for the College Entrance Examinations Board by the Educational Testing Service) were administered to college-bound secondary school seniors (Angoff, 1971). The task of ensuring that the scores reported for students taking one version of the SAT are comparable to those obtained by students taking another version involves *horizontal equating* of the different test forms. Similar procedures are used for state achievement tests, and other forms of assessment.

However, the Elementary and Secondary Education Act (often referred to as 'No Child Left Behind' Act) requires states to test reading and mathematics annually in grades 3 to 8, and at least once in grades 10 to 12 by school year 2005-2006, and science must be assessed at least once in each of the elementary, middle, and high school years by 2007-2008. The Act further requires that these tests be aligned with state standards and that the results be comparable from year to year (i.e. be horizontally equated). However, there is no requirement that the tests used in each of the grades be related to each other in any coherent way, and there is already evidence that, at least in some states, the tests have not been calibrated in a coherent way.

In order to investigate the extent to which the standards applied at the different grades within a state were internally consistent, Kingsbury et al (2003) compared the performance of students on state tests with vertically equated reference tests in reading and mathematics aligned to that state's content standards. They found, for example, that in Arizona, the proficiency standard set by the state for mathematics in the third grade equated to the 46th percentile of achievement on the reference test. However, the standard set by the state for proficiency in mathematics at eighth grade was equivalent to the 75th percentile on the reference test. This means that students at (say) the 50th percentile of achievement would be likely to be regarded as proficient in the third grade, but by the eighth grade, they would be regarded as well below proficiency, even though their progress was 'normal'. Of course, as Kingsbury et al note, there is no way of determining where each the standard for each grade 'should' be pitched, but these kinds of analyses can indicate where the standards are internally inconsistent. This is important, because the failure to equate tests vertically could lead to serious misdirection of resources. In the Arizona example, looking at the proportion of students achieving proficiency would indicate that the performance of middle schools was less good than that of elementary schools. The 'logical' response to this might be to concentrate professional development on middle school teachers, when in fact, no such conclusion would be warranted from the data.

A key priority for the design of assessment systems, therefore, is the need to ensure that the assessments are vertically equated, so that inferences about the progress made by students are warranted.

The issue of vertical equating is concerned with ensuring that the mean attainment of cohorts of students is measured in a coherent way. Once this is done the issue of the *variability* in achievement becomes important.

Growth models

In 1835, Quetelet noted that as the heights of Belgian boys and girls increased from birth to adulthood, so did the range of heights within a cohort of students of the same age, and the correlation of means with standard deviations appears to be almost universal, at least for physical measurements. For cognitive measurements, the picture is more complex. Time-indexed measures (e.g. by grade or age) do show much the same picture (see for example, Williamson et al., 1991; Wiliam, 1992). On the other hand, it has been found that some order-preserving measures (such as those that use item-response modeling) do not (Yen, 1986). Part of the reason for this is that item-response models used in most of these studies have fitted models to populations grade-by-grade, which has the effect of constraining the spread within each grade. In general, while item-response models do show lower rates of increase of spread, they do also generally show some increase (although there are some ceiling effects). However, the most important finding of this research is that there is no ‘natural’ way of measuring the growth of achievement over time, and that different models have different properties.

The choice of growth models is important, because it has profound implications for the distribution of attainment within a cohort, and the rate at which achievement increases over time. To see why, consider figure 1. It shows a hypothetical distribution of achievement across cohorts aged from 4 to 16 years of age with achievement being measured in “attainment age”. A 14-year-old student would have an attainment age of 12 if his or her achievement were equivalent to that of the average 12-year-old. Of course, whether there are such 14-year-olds is an empirical question, and depends on a range of factors, including the extent to which the domain can be regarded as unidimensional. The model shown in figure 1 assumes that the standard deviation of achievement within a cohort is one-tenth of the chronological age.

The importance of the relationship between standard deviation and age becomes apparent if we look at figure 2, which shows the distribution of achievement if the standard deviation of achievement is one-fifth of the chronological age. The peak of the distribution remains in the same place of course, but the amount of overlap between cohorts is much greater in figure 2 than in figure 1, and greater again in figure 3, which shows what happens when the spread is one-fourth the chronological age. At first sight, this may seem like a relatively arcane issue, but the rate of increase of spread within the cohort has profound implications for the trajectories of individuals, not least in terms of how many students are required to repeat grades (see below).

Figure 1: hypothetical distribution of achievement with standard deviation set to one-tenth of the chronological age

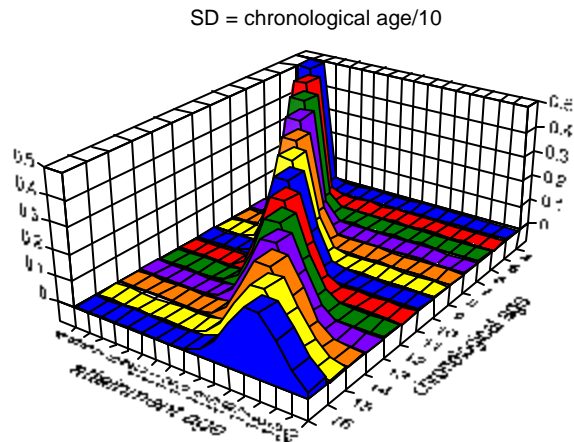
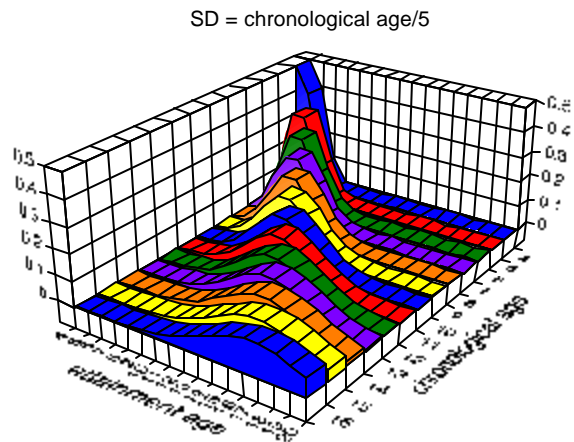


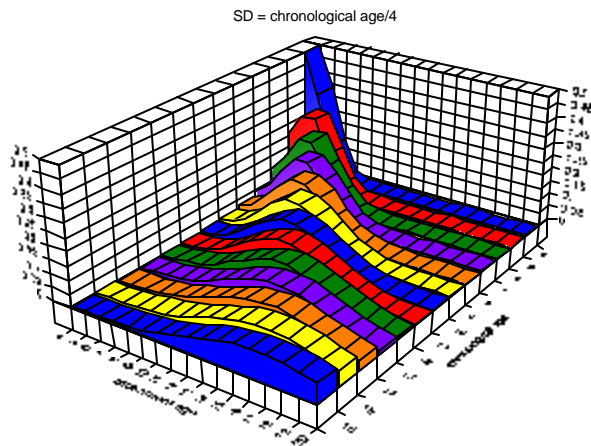
Figure 2: hypothetical distribution of achievement with standard deviation set to one-fifth of the chronological age



As well as being important, the spread of achievement within the cohort is a 'design issue', because different choices about the nature of the subject will produce different outcomes. For example, when ability in science is defined in terms of scientific reasoning, perhaps best exemplified by the science reasoning tasks developed by Shayer and Adey (1981), achievement will be less closely tied to age and curriculum exposure, and more closely related to measures of general intelligence. In contrast, when science is defined in terms of knowledge of facts that are taught in school, then opportunity to learn will be the most important factor—those students who have been taught the facts will know them, and those who have not will, in all probability, not. A third case might arise in the discussion of ethical and moral dimensions of science, where maturity, rather than

general intelligence or curriculum exposure, might be the most important factor. The crucial point here is that the spread of achievement, and whether the spread increases with age, is a consequence of the decisions taken in the design of the science curriculum and its assessment.

Figure 3: hypothetical distribution of achievement with standard deviation set to one-fourth of the chronological age



This manifests itself most tangibly in the relationship between the difference between cohorts of different ages and the range of achievement within a cohort of a given age. For the reasons noted above, there is no natural way of measuring learning, but if, for the sake of illustration, we take the amount that the average student learns in a year as a ‘measuring stick’ for learning, the range of achievement within any one class can be quite large.

For example, Brown *et al.* (1995) found that in England the performance of some students in the conceptual issues involved in measuring length and weight (mass) in grade 1 was well in advance of *most* students in grade 7, suggesting that there may be a ‘twelve-year gap’ between the weakest and the strongest in a grade 7 science class, which is consistent with similarly-focused research in mathematics (Brown, 1992 p. 12). Moreover, this is not a new phenomenon. The Sequential Tests of Educational Progress developed by ETS in the 1950s found that there was a ten or eleven-year gap between the strongest and weakest students in an age-cohort in listening, reading, writing, social studies, mathematics and science (Educational Testing Service Cooperative Test Division, 1957). Evidence from international comparisons suggests that the spread of achievement in the USA is amongst the greatest of all countries studied in both Mathematics (Mullis *et al.* 2000) and Science (Martin *et al.* 2000), although this is almost certainly due to a range of factors, rather than just the nature of the curriculum. Even in Japan, where the range in science achievement is amongst the smallest of the countries participating in TIMSS, the range is still considerable. Different countries have a range of methods of coping with this diversity of achievement within the age-cohort. Some countries attempt as far as possible to have all students doing the same science. In some

of these (e.g. Japan, Denmark) the students work through these together, while in others, all students work towards the same goals but some proceed faster than others, with the students being retained in grades until they reach the specified level of achievement (e.g. France, Germany) or allowing students to be 'socially promoted' but then grouped by ability within the cohort (e.g. England). In other countries, students of the same age work towards different goals. This can be through a common core curriculum which is supplemented by more demanding material for the ablest students (e.g. Belgium), by the designation of different 'streams' with different curricula within the same school (e.g. Netherlands) or by the creation of different schools with different curricula (e.g. Germany).

In the United States, there appears to be an uneasy compromise between these methods. Almost all of the states specify science curricula for each grade or group of grades. In the past, schools made substantial use of grade retention as a way of reducing the range of achievement within a class but its use declined in the 1980s and 1990s, and this may explain the wide spread of achievement in age-cohorts. However, many states are now considering making greater use of retention and ending 'social promotion', and this may be reinforced by the requirements of the No Child Left Behind Act of 2001. The act requires all states, districts and schools to make "adequate yearly progress" towards the goal of having all students assessed as 'proficient' in grades 3 through 8 by 2014. Because the definition of grade is by reference to the curriculum, rather than to students, there is no requirement for students to make progress, only that the school, district or state increase year-on-year the proportion of students in a grade assessed as proficient. Since failure to make adequate yearly progress can result in substantial sanctions for schools, up to and including the termination of staff and the reconstitution of the school, there will be a substantial incentive to achieve adequate yearly progress, and one obvious way to do this would be to retain students in a grade if it appears that they are likely not to achieve proficiency in the *next* grade the *following* year. The proportion of the cohort likely to be affected by such considerations depends critically on the spread of achievement within the cohort, and this in turn is determined by the way that the subject is defined.

Sex differences

Discussion of the causes of sex-differences in science achievement is beyond the scope of this paper. Here we concentrate on the size of the differences, and, more importantly, on their variability, as evidence of the mutability of these differences.

Willingham and Cole (1997) reported the size of sex-differences for seven science test batteries. The average sex-difference across the seven batteries was 0.17 standard deviations in favor of males, which is consistent with a range of studies (and very similar to the raw difference of 19 points difference between males and females in the US in the 1999 TIMSS study). However, what is interesting is that different test batteries give very different results. On the Armed Services Vocational Aptitude Battery (ASVAB) males out-perform females by 0.36 standard deviations, but on the 'Analysis of Science Materials' battery in the Iowa Test of Educational Development (ITED), females

outperform males by 0.17 standard deviations. Part of the reason for this is undoubtedly that the ITED requires students to interpret materials, placing a premium on reading comprehension and verbal reasoning, but this does not invalidate the finding. The important point is that the construct of 'science' can be defined in different ways, in some of which males out-perform females, and in others females out-perform males. This is a 'design issue' in that the definition of the construct of science has implications in terms of the size of any sex-differences that are likely to be found. It would, of course, be possible to define the construct of science in a way that resulted in equal performance of males and females: when Binet encountered gender differences in developing the first intelligence tests, he adjusted the selection of items to eliminate any overall difference. However, we are not advocating that this should be done for several reasons. First, 'fixing' the definition of science to eliminate sex-differences would be a 'blind' process that would run the danger of considerably misrepresenting the construct of science. Second, even if science were defined to eliminate sex-differences at a point in time, this would be unlikely to hold for any sustained period. Indeed, one of the strongest pieces of evidence that sex-differences in mathematics were not primarily genetic in origin was the finding that the size of the sex-differences in mathematics achievement in the United States had halved in the last fifty years (Feingold, 1988; Friedman, 1989; Hyde, Fennema and Lamon, 1990; Linn, 1992). What we do suggest is that the definition of science should be informed by the possible consequences in terms of sex-differences. Where a particular definition of science appears likely to produce considerable sex-differences, we do think it is worth exploring whether alternative, and equally legitimate definitions of science might produce different outcomes.

3. The locus of assessment

The terms 'classroom assessment' and 'formative assessment' are often used synonymously, but the fact that an assessment happens in the classroom, as opposed to elsewhere, says very little, either about the nature of the assessment, or about the functions that it can serve. Classroom assessments may provide a sound basis for summative assessments, and those conducted outside the classroom may provide valuable insights into how to take learning forward. As well as the locus of the assessment, we think that it is also important to attend to the issues of authority, resources, interactivity and scoring. Each of these is discussed in turn below.

Authority. The assessments may be generated by the teacher, or by outside agencies, or, somewhere between these two extremes. In many of Germany's *Länder*, assessments are proposed by the teacher, and approved by an external agency, such as a regional inspector. In Swedish upper-secondary schools, teachers are *expected* to modify the national tests before administering them to their students, if there are certain items that they feel their students will not be able to attempt successfully because the relevant subject matter has not been covered. Whether it is fair to assess students in different regions, or even students in different schools in the same region, on a different basis is, of course, problematic (see below).

Resources. The conditions under which students respond can be more or less controlled. At one extreme, typified by the traditional written examination, students may be required to respond alone, and without any additional materials. In other assessments, they may be able to consult specified textual resources (as in an ‘open-book’ examination), or a wider range of materials (for example, the internet) and even, in group projects, other students. In Denmark, for example, in the school-leaving examinations in mathematics taken at age 16, students are expected to work together to solve a problem.

Interactivity. In the traditional test or examination, there is a stimulus, to which the student makes a response, which is then judged. There is no scope for the student to ask for clarification of the meaning of the stimulus, and the rater is required to make a judgment of the response as it stands. In this context, it should be noted that the majority of classroom tests or ‘quizzes’ that teachers employ in their classrooms, as part of their normal classroom practice, are of this sort. In an oral examination, however, the student can seek clarification of the meaning of the stimulus, and the rater can ask the student to clarify or elucidate their response. Furthermore, the oral examinations allow the exploration of issues in depth and in some examination systems (notably Russia), the ability to interact with candidates is regarded as essential for valid assessment. Of course, while face-to-face oral examinations have been the traditional way of providing for interaction between student and rater, modern technologies allow a much greater range of options, including the possibility of having computers, rather than humans, conducting the assessments. Nevertheless, an inevitable feature of such assessments is that all students are not examined on the same basis, raising issues of fairness and equity.

Scoring: Where the results from the assessments are expected to serve summative or evaluative purposes, it is essential that the grades, marks or scores awarded depend as little as possible on who is doing the assessment—in other words that the assessments are *objective*. This is generally achieved by the use of machines, or by employing human scorers who have no knowledge of the student. At the other extreme, where assessments are intended to serve only a formative function, consistency of meanings across different raters is less important. What is more important is whether the assessments lead to improved instruction. However, just as teachers can author assessments for summative purposes, they can also be involved in the scoring of their own students’ work for summative purposes. One way that the necessary consistency of scoring across teachers has been achieved in the past is through scrutiny of the judgments made by assessors, which amounts to a kind of quality control process. Marks, grades or scores are generated, and at the end of the process, the quality of the assessing is inspected, and, if necessary, adjusted (this process is frequently termed ‘moderation’). What is important here is that while the assessment may be conducted by the teacher, it is done in a way that is inter-subjective—relying on the shared understanding of a community of teachers—so that the judgments are objective, in the sense that they are free from individual subjectivity. Furthermore, the assessment is carried out against a set of standards that are determined by the community rather than the teacher, so that even though the teacher is involved in scoring the student’s work, the teacher is, in a very real sense, the student’s ally rather than their enemy (e.g. “I’d love to give you an A for this but you just haven’t reached the required standard yet”).

The assessment systems in Australia (Queensland), New Zealand and Sweden signal a move away from a traditional quality *control* orientation towards one of quality *assurance*. The major effort goes not into correcting scores assigned by teachers, but into improving the ability of the teachers to get it right first time. In many systems is also noteworthy that the focus is on securing consensus not through getting teachers to agree on some lowest common denominator, but through beginning to address explicitly the features that are likely to be present in good responses. The notion of ‘community of practice’ (Lave and Wenger, 1991) is a useful idea for thinking about how teachers can come to consensus over the marks, grades or scores to be awarded to students’ work, but it can also serve to disguise what it is that they come to agree *on*. After all, the requirements of reliability are met if teachers’ judgments are consistent, even if they have no idea what they are doing, or how they are doing it. The result of this can often be that teachers can judge accurately the standard of students’ work, but have little idea about how to improve it.

In contrast, getting teachers to meet together to talk about what makes for high-quality work in science not only improves the consistency of the judgments they make of students’ work, but also provides a valuable form of teacher professional development in its own right.

4. Extensiveness of assessment

A key theme running through several of the issues raised in the previous section is the extent to which all students are assessed on the same basis. Traditional wisdom dictates that fair assessment can be attained only if all students are assessed on the same basis, and this is the notion of ‘fairness’ used in traditional tests. However, in the same breath, it is also routinely acknowledged that it is essential to make adjustments to assessments for particular populations, such as students with visual impairments, specific learning disabilities (such as dyslexia) or motor impairments. At the higher levels of the educational system, it is routinely accepted that at least part the purpose of the assessment is to provide candidates with an opportunity to show what they can do, through the use of non-uniform assessments such as coursework, projects and theses. Of course, it could be argued that the requirement for non-uniform assessment at these higher levels arises from the complexity of the judgments that are necessary, but then the same also applies to earlier stages of the learning process—recent research has shown convincingly that the state of anyone’s learning is a complex schema which defies simplistic analysis. Failure to recognize this (or, perhaps even worse, recognizing it but failing to acknowledge its importance) has resulted in a simplistic approach to assessment that leads to an emphasis on low-level aims that weakens validity. If more complex notions of fairness than simply making sure that all students are asked the same question are felt to be necessary for certain populations, then why not for all?

The No Child Left Behind Act requires that the assessment system in place in each state delivers a determination of the level of achievement for each student as ‘below basic’, ‘basic’ or ‘proficient’, and these are aggregated to provide indications about whether specified sub-groups in the school are making “adequate yearly progress”. However, the

Act also allows students with special needs to be assessed on an alternative basis, which takes into account their special needs. While the number of students assessed on an alternative basis that can be regarded as 'proficient' for the purposes of determining whether the school is making adequate yearly progress is limited (currently 1%), the report to the student's parents is the level of proficiency achieved on the alternative assessment. This is an important provision, since it means that it is not possible to aggregate the scores reported on individuals to provide the outcome used for accountability. In other words, there is no direct link from the way that outcomes are reported for summative purposes to the outcomes reported for evaluative purposes. This differs from the situation in, for example, England, where it is possible to aggregate directly from the points earned by individual students on each item on a test paper to the position of the school attended by those students in the national 'league table' of performance.

This in turn raises issues of the extensiveness of the assessment, and in particular, whether all students need to be assessed on the same basis. As noted above, giving all students the same test at the same time provides an apparently fair basis for comparison, but has two major drawbacks. The first is that using timed written tests weakens the reliability of the assessment, because only a small sample of the content for the grade can be assessed in such a test. The reliability is weakened because of the variability of student performance (students have 'good' days and 'bad' days) and because of the particular sample of items that is selected for the test. The second drawback is that only a limited proportion of the standards specified for science in a given state can be adequately assessed in timed written tests, and teachers can predict which kinds of standards will, and will not, be assessed (Wiliam, 1993).

These concerns have led to the use of 'matrix sampling' models for national evaluations of performance such as the National Assessment of Educational Progress (NAEP) in the USA, the now defunct Assessment of Performance Unit (APU) in the United Kingdom, and the new national monitoring arrangements in New Zealand. However, these surveys are almost always quite independent of assessments for the certification of individuals, or for holding individual schools to account, and rarely connect in any meaningful way with learning.

5. Assessment formats

There is a wide range of methods available for assessment: the choice has to be matched to the purposes and to constraints on time and cost. The main possibilities are as follows.

Fixed response questions have many advantages, notably in coverage and reliability of scoring, but can have bad feedback effects on learning habits. The main disadvantages are that there is no direct evidence of students' reasons for their choices, so their value for formative and diagnostic purposes can be limited (unless the distractors are based on well-known misconceptions, which currently appears difficult and expensive to do). Other disadvantages are that the knowledge or reasoning that is tested will be in an isolated or restricted context, and that high-stakes regimes with teaching to the test lead

to an atomized approach and a passivity in which learners' judge other people's ideas but do not propose, formulate or create their own ideas.

Open response questions give more information about students' thinking and can still take little time and use well controlled marking procedures. There is a great variety of such formats. Some common examples are:

- propose words to fill the blank spaces in a set of sentences;
- solve a numerical problem which requires only a small number of steps;
- require students to attempt a complex problem through tackling a guided sequence of component steps;
- on the basis of a supplied text or set of data, answer a set of short questions designed to test understanding of the text, or skill in responding to and handling new evidence.

Such questions can be used to assess knowledge, reasoning and skills at various levels of complexity. They retain some of the advantages of fixed response questions in that a fair number can still be tackled in a short time, and marking can still be reliable. However, as Gauld (1980) found, students may often misread such questions, and thereby fail to do themselves justice.

Essay questions are unique in exploring complex structures of knowledge and reasoning; the intended demands have to be specified to students in some detail, and the problems of reliable marking have to be tackled, in particular by use of multiple marking of each response. Compare, for example, the question:

Write an essay on the applications of physics

with the following alternative:

Write an essay on the applications of physics. You should choose only one of the three main areas of physics you have studied this term—waves, dynamics, electromagnetism. For the one that you choose, describe three applications, explaining for each how it is to be understood in terms of the principles of physics and why it is of practical importance.

This second example sets out the material on which the essay is to be based, specifies the kinds of reasoning needed, and conveys ideas about the criteria that will be used in the scoring. The short version would probably evoke a wide range of answers, and the task of marking such variety of answers in a fair way would be almost impossible.

Certain verbs which are commonly used in essay questions—such as identify, describe, state, explain, compare and contrast are often ambiguous, and their use can sometimes hide a vagueness in the examiner's own mind.

For the marking of essay questions many research studies have explored the differences between analytic or 'checklist' marking, and holist or 'impression' marking. There is no clear outcome in favor of the one or the other, but it does seem clear that it is better to have two rapid markings by two independent examiners than one slower exercise by one person alone (Wood, 1991). However, because the number of such essays that can be attempted in the time available for assessment is small, there is a significant problem of task-student interaction—in other words, some students get high scores because the questions they were asked were ones that particularly suited them. Since the most expensive aspect of such assessments is the cost of scoring, studies have explored whether it is better to have a number of tasks scored twice, or twice as many tasks scored once each. In terms of reliability, the answer is unequivocal: it is better to have twice as many tasks scored once each (Linn and Baker, 1996), although it may be difficult to persuade those who do not understand the psychometrics that this is, indeed 'fairer'.

Essays for summative assessment purposes may be written under examination conditions, or may be written on the basis of library research, with free access to books and notes. Possibilities of this type overlap into the field of performance assessment.

Performance and authentic assessments can be composed of a variety of classroom tasks and automatically meet some of the requirements for validity. However, the procedures for ensuring reliability in assessment across different teachers and tasks require careful attention.

It is difficult to give a precise definition to delimit the activities described under these headings. One unifying idea is that it is to do with assessment of activities that can be direct models of the reality to be assessed rather than disconnected fragments or surrogates. As Airasian puts it "Rather than asking students to *tell* what they would do, performance assessment requires that they *show* what they can do" (Airasian, 1991 p. 252).

Much of the US literature speaks both of 'performance assessment' and of 'authentic assessment' and these two are sometimes used interchangeably. Darling-Hammond et al. (1995, pp, 3-4) implies that distinctive characteristics of authentic assessment tasks are that tasks are embedded in the curriculum, that an assessment is made of students' response to a genuine learning experience, not a contrived one, and tasks are set in real contexts that connect school work to real world experience.

Classroom work often involves work in groups, and so authentic assessment requires that the group aspect of the work be explored. Where the group work is simply a means, the task is to disentangle the achievement of individuals from the group process, but where collaboration in groups is an end in itself, the contribution to the group process is an aspect to be assessed (details are given in Wood, 1991 ch.16).

A variety of approaches—*portfolio* assessment, graded assessment, modular assessment and records of achievement—have enriched the approaches to testing by breaking away from the dominance of the single terminal test and by promoting a widening of the range of students' characteristics that can be assessed and attested.

In a UK initiative that was given impetus under the title of Records of Achievement, the development was marked by tensions between the formative and summative philosophies. The formative emphasis implied that students could negotiate the record and come to agreement as to its contents and its judgments, but some found it hard to give students this degree of authority. The initiative suffered when it was side-lined by the onset of national testing (Broadfoot, 1986).

Portfolio programs have also attracted attention and investment in the US. One of the motivations has been to set up state systems based on classroom assessment, both to improve school's reports and graduation procedures for individuals and to complement or even replace the state systems, based on external standardized tests. Evaluation studies have revealed that, in spite of the extra work involved, most teachers welcomed the approach. However, it was also revealed that the reliability of the marking was low, with unsatisfactory agreement between different markers and large variation in the performances of a student from one task to another. Thus, it seemed the training of the teachers in marking, and the extent of work assessed, might have to be extended, calling into question the cost (Koretz et al., 1994) and the procedures tightened in ways that would reduce the both validity and teachers' commitment (Stecher, 1998). Other similar experiments, notably in Kentucky and Pittsburgh, have also encountered such difficulties.

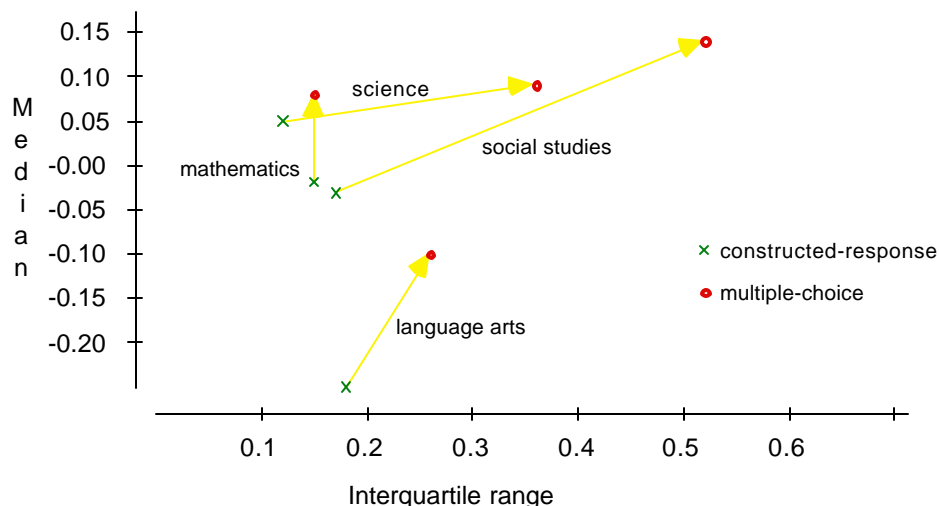
Overall, both the reliability and validity of a test result can be enhanced if cost and time can permit the deployment of a range of methods, so that the results that they supply can be combined. The choice of tasks is a critical feature if the learning and assessment aims are to be attained. Here constraints of cost and of testing time can give rise to severe limitations. Fixed response tests are relatively inexpensive and US opinion seems to balk at the costs of more 'authentic' methods, even although such costs are accepted on a routine basis in many other countries. The particular problem for performance or authentic assessments is that variability between different tasks contexts makes it hard to achieve high reliability in any approach that does not make use of results of normal classroom work over a length of time.

Sex-differences by assessment format

Ryan and DeMark (2002) collected data on the size of male-female differences in tests of different subjects, according to whether the tests used multiple-choice or constructed-response items. Consistent with the research literature on this issue, they found that for a given subject, the multiple-choice tests showed a bigger male advantage than constructed-response tests. A move from constructed-response tests to multiple-choice tests might therefore be expected to increase the scores of males relative to females. However, what they also found was that the sex-differences in multiple-choice tests were much less variable than those found for constructed-response tests. So if we imagine moving from constructed response items to multiple choice items in science, the median effect size in favour of boys would increase modestly from 0.05 standard deviations to around 0.07 standard deviations. However, the inter-quartile range of the effect-sizes increases from 0.11 for constructed-response tasks to 0.35 for multiple-choice tests. This suggests that, at least as far as sex-differences are concerned, item format is less

important than effects of task dependence within a given format— so that there are implications for task selection and so for the way that science is defined.

Figure 4: Median and inter-quartile range of sex-differences by item format



6. Scoring models

Achievement in science is multidimensional. Very few students are equally good at biology, chemistry, earth sciences and physics, and even within a domain, each student will have relative strengths and weaknesses. Some may have a good grasp of concepts, but be less good at recalling specialized vocabulary. Others may be stronger on processes of inquiry and less good on the subject matter. Despite these well-known variations, we have to collapse all the fine-grained information we have on individual students to produce descriptions of individuals such as ‘proficient’ or ‘basic’, and even these summaries of individual achievement are themselves aggregated to determine whether schools are making ‘adequate yearly progress’ or not.

The way that these processes of aggregation are carried out have profound implications for the extent to which summary judgments about individuals or institutions can be ‘reverse engineered’ to provide information about the meanings of those judgments. For example, the aggregation rules specified for determining whether a school is making ‘adequate yearly progress’ (AYP) as defined under the No Child Left Behind Act mean that if the school is determined as making AYP, we can be sure that each sub-group of students is making good progress. However, by the same token, if one sub-group fails to make progress, then the whole school is deemed to be failing to make AYP. Rules such as those that are used to determine whether schools are making AYP are sometimes referred to as ‘conjunctive’ aggregation rules. The strength of such rules is that they allow strong conclusions to be drawn—if the school is making AYP, then each sub-group

is making progress. Their weakness is that schools where a single sub-group is just failing to make sufficient progress are lumped in with those schools where all students are failing to progress, because no ‘compensation’ between groups is permitted (in other words, strength in the performance of one group is not allowed to offset weaknesses in another).

For the assessment of individual students, such conjunctive aggregation rules are rarely used. When a student is assessed as ‘proficient’, we do not generally know what they can do, because we usually allow weaknesses in one aspect of the subject to be offset by strengths in another. A good example of this is provided by the assessment of the mathematical competence of student nurses, where it is frequently assumed that a particular score on a mathematics test confers a guarantee that the individual is skilled on particular kinds of calculations that are in the specifications for the test, but in fact, such a conclusion is frequently not warranted, especially on topics that students find difficult, such as ratio (Pirie, 1987).

In all this, it is important to bear in mind that the score required in order to be regarded as ‘proficient’ is arbitrary in two senses. In the first place, what it means to be ‘proficient’—i.e. the performance standard—is arbitrary. Analysis of NAEP data and their relationship to test scores indicates that there is one state whose standard for grade 8 is equivalent to another’s standard for grade 4 (Braun, 2004). Secondly, even once we have agreed on the performance standard, the score that equates to this is still a design choice—in other words, the cut-score that equates to a particular performance standard is also arbitrary. We could set a moderately demanding test, where a score of 50% on the test would be equivalent to the performance standard, or we could set a much easier test, where a score of 80% would be required to provide evidence of the same standard of achievement.

Setting a test so that a given performance standard requires getting most of the items correct has two major advantages. First, we know that students who are judged ‘proficient’ have managed to answer most of the questions correctly, and we can therefore draw reasonably strong conclusions about what the student can do, or at least could do at the time of the assessment (although it is worth noting that even if a score of 80% of the available points is required for mastery, a particular student may have dropped all their points in one area). Second, the experience of most students taking the test will be positive, in that they will feel that they have ‘done well’.

However, setting such a high cut-score will also have some unfortunate consequences. The reliability of the classification of students as ‘proficient’ or ‘not proficient’ is maximized when all the items are pitched at the borderline between the two categories. If the test is very easy, then the decision about whether a student is proficient or not is made on the basis of a small number of items, and so the consequences of a small error are magnified. We are, in effect using only a small part of the test to make decisions, which means that our decisions are less reliable than they could be for the same amount of testing time (or to put it another way, we are spending a lot more money on our tests than we need to, given the reliability we are getting).

This then presents us with another ‘design choice’. We can construct tests to have high cut-scores, leading to positive experiences for students and the ability to ‘reverse-engineer’ reported outcomes, but with less reliable classification. Or we could maximize the reliability of the decisions, by pitching the items at the performance standard, but then many students will find the test hard, and when we know that a student is proficient, we won’t know what they are proficient *on*.

The issue of compensation between different components of an assessment also has consequences in terms of teachers’ practice. Teachers who are preparing their students for high-stakes tests routinely engage in a form of cost-benefit analysis in which they trade off the benefits of teaching particular topics against the cost in terms of classroom time. The result is that parts of the curriculum that have a low probability of being tested have a low probability of being taught. Where a compensatory model, such as total number correct, is used for scoring, teachers are more likely to decide that some topics are just not worth the time needed to teach them effectively.

Of course, most scoring models are neither completely conjunctive nor completely compensatory, and there is considerable scope for designing scoring models that incorporate incentives to promoting balanced growth in teaching. For example, the original version of the national curriculum for Technology in England and Wales reported student achievement on a profile of four aspects of technological capability: identifying the need for technology, problem specification, solution, and evaluation, each to be reported on an eight point scale. Because each of these four aspects of capability were regarded as essential, the overall reported level was to be the highest level achieved on three of the four aspects, provided the fourth was no more than one level below the reported level. The same idea could be applied to the requirements of the NCLB Act by (for example) requiring that a student would be deemed proficient only if they were proficient in three out of four of the sub-domains in a subject, *and* they achieved at least a ‘basic’ level of achievement in the fourth. With such a scoring model, the teacher’s attention would shift according to the pattern of students’ strengths and weaknesses. Unlike traditional scoring models, emphasizing one aspect at the expense of others would not benefit most students.

The major drawback of such a scoring model is that the reliability of the reported outcome is effectively that of the weakest component score (Cresswell, 1994). One way of avoiding this would be to derive the overall outcome as before, but to report profile score separately. While the separate profile scores would not be as reliable as the overall score for individuals, for groups of students they would be quite stable, and thus would provide a useful tool for schools, districts and states to monitor the extent to which different aspects of science were being taught.

It is also important to pay attention to the way that a subject is broken up into components. Profiles of performance could be presented in terms of ‘traditional’ divisions such as biology, chemistry, earth sciences and physics, or by schemes based on the nature of scientific activity, focusing on, for example, interpretation of existing data, planning and carrying out experiments, making observations, representing data, and drawing inferences from data. Such an approach would emphasize the links between

similar aspects of the sciences, and would help reduce the extent to which students see the separate sciences as unrelated.

The aim of the No Child Left Behind Act was to force schools to become ‘high reliability organizations’ in which failure in some aspects could not be compensated by success elsewhere. The choice of scoring models used at the student level may have equally profound consequences for the pattern of strengths and weaknesses within students.

7. Quality in assessment and testing

There are two main criteria of quality of an examination result that should be a basis users to have confidence in its results are: *reliability* and *validity*.

Reliability

When a test is not *reliable*, the score that someone actually gets on a particular occasion will not reflect their capability. However if they took the same, or similar, tests on a number of occasions, then the average of all those scores *would*, in general, be a good indicator of their capability, and this long-term average is sometimes called the ‘true score’. The crucial relationship is therefore how close the score we get on a particular testing occasion is to the ‘true score’. If we take a test with a reliability of 95% (a reliability generally only achieved with specialized psychological tests), someone with a true score of 50% will, on the majority of occasions, get a score between 47% and 53%, although of course, the score on any one occasion could be outside this range. With an 85% reliable test (the reliability typically achieved with educational assessments), the band would be wider—between 45% and 55%. Given the possible error in a final mark, there follows a possibility that a candidate’s grade, which is based on an interpretation of that mark, will also be in error.

Thus this criterion of *reliability* is concerned with the inevitable chance of error in any test of examination result. The main sources of error that can threaten the reliability of an examination result are:

- Any particular student may perform better or worse depending on the actual questions chosen for the paper that year;
- The same student may perform better or worse from day-to-day;
- Different markers may give different marks to the same piece of work.

The first of these three is a problem of *question sampling*. Questions can differ both in the content, and in the type of attainment that they test (e.g. knowledge of definitions, solution of routine short problems, design of an experiment to test a hypothesis). Examiners employ systematic ways of sampling the content domain, and may deploy different types of questions, choosing between multiple-choice question to test knowledge and simple applications to sample any content areas, and small number of longer problems to test application of concepts and

synthesis of ideas (e.g. design of an experiment involving detection of light with devices which give out electrical signals).

Thus the composition of an examination is a delicate balancing act. The shorter the time allowed for the test the smaller the sample, and the smaller the confidence one can have that the result for any one candidate is independent of the particular sample attempted. Any examination can become more reliable if it can be given a longer time.

The size of the errors due to the first of the sources of error listed above can be estimated from the internal consistency of a test's results. However, if checks on internal consistency reveal (say) that the reliability of a test is at the level of 85%, then in order to increase it to 95% it would be necessary to more than triple the length of the test. Indices based on such checks are often claimed to give the reliability of the examination result: such a claim is not justified for it takes no account of other possible sources of error.

The second source of error means that the actual score achieved by a candidate on a given day could vary substantially from day to day. Again, this figure could be improved, but only by setting the test in sections each taken on different days. Data on this source are hard to find so it is not possible to estimate its effect: it would seem hard to claim *a priori* that it is negligible.

Where questions demand constructed responses, the marker error is also a problem. In countries where most or all of high-stakes tests use constructed response questions, particular cases of marker error justifiably attract public concern, yet overall, the errors due to this source are probably negligible in comparison with the effects of the other sources.

Overall, the main limitations on the accuracy of examination results could be offset if increases in costs, examining times, and times taken to produce results were to be accepted by the educational system. Such acceptance seems most unlikely: in this, as in many other situations, the public gets what it is prepared to pay for.

Whilst 'reliability', meaning internal consistency, is commonly quoted for fixed response tests, it is rarely researched for other types of public examinations. Yet there are a few such, by Rogosa for state standardised tests in California (Rogosa, 1999), and for high-stakes constructed response tests used in the UK, by Wiliam and Black (1996) and notably, Gardner and Cowan (2000). These all show that proportions of at least 30% of candidates are wrongly graded. The effects of mark errors on grades are often large because grade boundaries are set only a few marks apart in an attempt to differentiate around the peak of the mark distribution.

Comment [DW1]: Page: 27
I don't think I have a copy of this—I assume it's what we wrote for that meeting with the DfES people.
Deleted:

Validity

Traditionally, this is taken to be the extent to which a test measures what it is supposed to measure. It can be compromised by attempts to enhance reliability. For example, if the domains to be sampled, of question content and question type, are narrowed, the sampling error is reduced, internal consistency of a test would be increased, but the information the results convey to a user would be changed. The potential for harm of such a change is brought out by Messick's (1989) definition of validity, which is that it concerns the adequacy and appropriateness of the *inferences and actions* based on the test results.

Another factor affecting validity bears on whether or not questions have been so composed and presented that the student's response will give an authentic picture of the capability being tested—a feature which may be called the *disclosure* of a question. Good disclosure is not easy to attain. For example, several research studies have established that in multiple-choice tests in science, many of those making a correct choice among the alternatives had made their selection on the basis of incorrect reasoning, whilst others had been led to a wrong choice by legitimate reasoning combined with unexpected interpretations of the question; it would seem that in such tests, about a third of students are incorrectly evaluated on any one question (Tamir, 1990; Towns & Robinson, 1993). It has also been shown, for open-ended questions, that mis-interpretation frequently leads candidates to fail to display what they know and understand.

A more difficult threat to validity is posed by the problems of bias in test papers (or more precisely, bias in the inferences that are made on the basis of test results). Many studies have shown evidence of systematic differences in the meanings of test results for students of different gender, or from different socio-economic groups or from different ethnic backgrounds. For example, it is found in many countries that males outscore females in physics. At one level, such a finding is unproblematic, in that we can safely conclude that the males performed better than the females on that selection of physics items provided our sample is large enough. But if we use such items to predict which students will go on to become the best scientists, there is a serious possibility of bias. Some attempts are made by test producers to reduce bias in individual items, but research evidence directly relevant to their testing is often inadequate, and there is a basic problem about whether or not to make adjustments on an *a priori* belief that all groups must be equal. The average sizes of such bias effects are usually quite small, but it is hard to be sure there might not be significant effects on particular individuals.

There are more subtle problems of validity involved when complex skills are being assessed. An example would be the ability to carry out an experimental investigation in a laboratory. Limitations of time and cost might make this so difficult to assess directly that examiners might decide to use a written substitute—e.g. asking students to write an account of how they would carry out such an investigation rather than asking them to do it with real equipment. In general such 'surrogate' methods are of questionable validity: this has been shown by asking the same students to make a written account, and then actually to carry out the task in a laboratory. There is usually very low correlation between the two performances, so that it can at least be said that any 'surrogate' would have to be tested empirically before it could be accepted as valid. Even if the correlation between the surrogate and the practical experiment were high, there is no guarantee that such correlations would hold up in the future. If teachers stopped conducting practical labs and concentrated on teaching students how to write accounts of experiments, then the 'surrogate' would cease to be a good predictor of actual performance in labs.

Understanding limitations

Research evidence about the limitations to the reliability and validity of examination results is needed, both for those who make policy for such systems and for the users of the results. For policy makers, the challenge is to optimize the overall strategy within the limited resources of expertise, time and money. The use, for example, of assessments made by teachers, is limited by

fears about their reliability by a public who are ignorant of the limitations of the tests that they trust instead.

Users should be aware of the inevitable limitation of test results in order that of they can use them appropriately in making important decisions. Professional standards for all tests require that data on these limitations be produced and published (AERA, 1999). Thus, if an employer or a higher education institution were to believe that a test score was free of all possible error, other evidence about an individual might be ignored, whereas if it was known that there was a significant margin of error for the reported scores, users of test information might well consider giving more weight to other information.

8. The role of teachers

The possibility of using teachers' assessments as the main or sole source for certification purpose is attractive because it offers the prospect of improving both reliability and validity and of securing better alignment between formative and summative assessments, to the advantage of both functions. The problems that make this seem a formidable target are the expertise of teachers and the difficulty of aligning and trusting their judgments. The importance of this issue calls for the special attention given to it in this section. The role of teachers in assessments has already featured in several sections above. The purpose of this section is to draw some of these threads together.

The system in Queensland is distinctive in that it presents perhaps the best solution (in terms of realizing the advantages and overcoming the problems of involving teachers in assessment) that we have found in operation on a large scale anywhere in the world. In the Queensland system, the summative results that are used by employers and by higher education depend solely on teacher assessment, which is based on student portfolios and a system of inter-school moderation. The building of this system took several years, but it has now been securely established for over thirty. Initially teachers needed a great deal of support. First, they needed help to break their initial reliance on the types of test that the earlier external test system used. Second, they needed to develop formative skills, in part so that they were better able to guide and judge students' use of portfolios. Third, they needed to develop the procedures and skill for the conduct of moderation meetings at which samples of student portfolios have to be exchanged between teachers to help arrive at agreement on common standards. The system of moderation is complex and rigorous. The assembly of assessment for the moderation process has to be related to the state's curriculum standards in each subject, and the first moderation meetings are to approve work programs a year before the final assessment. In the final stages, proposed results and samples are brought to meetings for cross-school discussion and confirmation—or adjustment. There is then a third stage in which selected samples are sent to the state's Board of Education for further scrutiny and confirmation.

The main rules about the contents of portfolios are that each must contain evidence pertaining to all of the state standards and corresponding criteria, and each piece of such evidence must be the latest and best evidence for the standards it reflects. The contents can be flexible and varied in form (e.g. test paper results, project reports). The content of

the portfolio can evolve, in that if a piece of work related to a particular standard improves on an earlier one, the older one is to be discarded. Thus assessment is meant to be progressive in allowing for improvement over time. The teacher can give help but must make allowance for such help in scoring the work. The variety of evidence allowed together with its range over time help secure a high level of reliability and validity. In guiding their students to improve particular work and giving advice as appropriate, teachers are helped to align their summative with their formative roles. The fact that teachers are given full summative responsibility is seen by the state authority as important: as one senior officer put it, "Teachers take up the challenge when they are given the responsibility" (Maxwell 2004).

The framework and rules for this system are one feature of its success. The other is the commitment of teachers. It has been found here, as in similar work in the UK, that the most effective way to help teachers internalize standards and refine their judgment is through the professional conversations that go on when they are trying to evaluate the concrete examples in the various portfolios that they exchange during moderation. Such meetings are thus occasion for professional development as well serving the assessment needs.

Recent evidence in the UK has shown the development of formative practice is no easy matter, because it involves fundamental changes both in the way that teachers understand learning and in the way that they relate to their students. Thus, whilst these changes can be achieved in an intensive program of professional development (Black et al. 2003), attempts to replicate these outcomes with less expensive methods of in-service training have had mixed success. Whilst most teachers welcome the encouragement to develop their formative work, many seem to believe that they have achieved good practice when they are far from doing so.

It seems clear, not surprisingly, that the tensions created by summative pressures is an important source of difficulty, at two levels (Black, Harrison and Hodgen 2004). At one level, pressure to 'cover' the content that high-stakes tests will examine tells against spending time on developing understanding, an effect which is exacerbated by the weak validity of test questions which do not call for thoughtful and constructive explanations. At a more subtle level, continual pressure to ensure that narrow test targets be met leads teachers to believe that once their students can recall the content and tackle exercises of low cognitive demand, their task is complete. If they seriously tackle the formative task, teachers own views of learning, often implicit and weakly theorized, are challenged.

Such problems would be more manageable if teachers had more control of and responsibility for high-stakes assessment. The conflicts and synergies between purposes would then play out, at least in part, within teachers' own ways of handling both summative and formative responsibilities. However this is done, it must be constrained, but perhaps enriched, by the need for a moderation process. One way to do this is to use external tests as a calibration device for scaling results from different teachers: this is inexpensive, but because it means that good results in the calibration tests are all important the external pressures are still there, whilst the classroom assessments can be given scant attention, as they will achieve value by the external test calibration rather

than by their own quality. However, this approach is used in Sweden without any reports on such negative effects.

A second method is to use *inspection*, either by a weak method of collecting samples of written work by pupils and having external examiners grade them to compare with teachers' grades, or by a stronger method by having 'moderators' visit schools, examine work of all kinds, and interview some of the pupils. This method is used in the UK for a contribution made to the summative results by teachers assessments of science investigations carried out in normal school contexts. The effects have been deplorable (Black et al. 2004), because the anxiety of teachers to conform to the rules and ensure 'good' results has resulted in stereotyped exercises that are a travesty of scientific investigation. A similar negative outcome was reported by Stecher (1998) who found that teachers' practices in the conduct of portfolio assessments were narrowed down to 'rubric-driven instruction' as the twin requirements of reliability and validity imposed constraints.

A more relaxed and radical form of this second approach is to inspect an institution's practices and evidence in detail, and once these have given satisfaction, license the institution to grant certificates for (say) a few years and leave it alone until the next time for inspection comes around – an approach which has long been used in the UK for vocational qualifications.

A third method is the use of *group moderation* meetings, as described above for Queensland. The experience there, following the detailed guidelines, shows how teachers' assessments, given the range and variety of assessment events, can achieve higher reliability than external tests and can also help achieve higher standards of validity in that the contexts for assessment reflect those of the learning experience, and resemble the contexts of future application of the learning more closely than the context of the examination room.

A basis in teachers' summative work also opens up the possibility of using a mastery learning approach in that progress within school can be guided by requiring high standards at one type of work before moving on to more demanding levels. Then the assessment result is characterized by the top-most level of mastery rather than by a means score across a range of levels, and a terminal summary test is irrelevant. A corollary however is that teaching plans must be framed around a well-founded model of progression in the learning of each discipline. The Queensland system incorporates some of these features. Such 'graded assessment' systems were operated, for a few years, in England in several subjects including science: they were very motivating for students, but was abandoned when government insistence on adding an overall terminal test reduced its attractions and added too much to the cost (Brown, 1998).

There are however some limitations. One is that plagiarism, variations in the help given to students with informal assignments, and personal bias of the teacher, can undermine fair and equitable judgments. There is also the issue of the quality of the assessment demands, for teachers cannot deploy, in the construction of assessment tasks, the expertise in either the subject discipline, learning theory or assessment technique, that an

external agency can deploy. These advantages can be overcome and there is extensive research literature on teacher assessments to guide the design of systems to do this (Harlen 2004) but, as the Queensland experience showed, significant investment to help teachers develop appropriate professional expertise would be needed in establishing any innovation which required a new reliance on teachers judgments.

Given all these considerations, there remain several design choices about the use of teacher summative assessments or certification and accountability purposes. One is about the ways in which teachers' assessments are produced, i.e. whether by a sequence of formal tests, by assessments of set pieces of written or investigation work, by observation of on-going activities, by assembly of such evidence within portfolios, and so on, plus the choice of how such evidence is aggregated to produce a final result. Another choice is for the role of teachers' results. They can be published alongside the results of external tests, or they can be combined with such results in a variety of ways, or they can be the main source of results subject to external calibration or to some process of moderation.

9. Contextual Issues

Cognition, Community and Policy

Learning Cognition and Motivation

The behaviorist model of learning comports well with a stimulus-response approach to assessment, and both of these ally well with 'teaching to the test' designed to 'deliver' and reinforce a curriculum packed with content. This scenario illustrates the close links between curriculum, pedagogy and assessment and their mutual dependence on theories of learning. In a behaviorist approach to learning, there is emphasis on recall, on lower-order thinking skills, whilst commitment to development of understanding is tempered by an assumption that this will develop later on the basis of remembered information. Such assumptions have strongly influenced educational practice, provoking the comment that current assessments arise from 20th century statistics applied to 19th century psychology. One sign of their deficiency is the evidence that high-stakes tests lead to short term improvements as teachers learn to teach to them, but that such improvements are illusory as they are to be found only in the context of responses to the tests which engendered them (Linn 2000). Another is that teachers who teach for understanding do produce better performance results on standard high-stakes test than teachers who teach directly for performance in these tests (Newmann et al. 2001).

Practice has still to take advantage of the many advances in learning theory and assessment technology that have displaced behaviorist views (Bransford et al. 1999, Pellegrino et al. 2001). New principles of learning, notably that structures incorporating effective associations are essential aspects of memory, that new learning has to be reconciled and integrated with pre-existing understanding, and that higher-order thinking skills develop from a quite young age, should all have a powerful effects on strategies for

teaching and learning. Two specific ideas bear directly on the practice of assessment. The first is that the context of new learning is all-important, so that generalizing across many contexts is problematic. The second is that it is not possible to use broad principles of learning to derive the sequences in which learners come to understand and use specific disciplines. To develop well grounded strategies for teaching (say) number operations or the reading of texts, it is necessary to conduct detailed empirical studies which can analyze the ways in which learners develop understanding in each area, so charting complex patterns of alternative routes and identifying common obstacles. Any assessment that is not informed by such research may yield no useful formative evidence in that the reasons for, and the significance of, the failures cannot be inferred from the results. A valid summative assessment may be less fragile, if users are only interested in the presence or absence of competence, but in any regime other than a mastery learning system, the interpretation of a score (say) of 50% to make inferences about future achievement in the domain is fraught with uncertainty. Thus sophisticated understanding of learning can lead to more stringent criteria for validity.

The development of well-grounded schemes to assess progression within specific domains of learning will draw heavily on domain-specific research. Given that patterns of learners' responses will inevitably be complex, it will follow that interpretation of the responses will require care, although the level of sophistication of the psychometrics will depend on the context of use – a national test and a single oral question in a classroom are quite different in this respect. What is also clear (e.g. from Pellegrino et al. 2001) is that there are few adequate combinations of research into learning progression with well-grounded sequences of assessment tools, and even fewer examples of such findings being put into practice on a large scale.

Another aspect that cannot be ignored is the effects of teacher-student interactions on the confidence (in learning) and the motivation of students. Where high-stakes tests so oppress teachers that their pressures filter down to their students, negative effects on the motivation and confidence of the students will follow (ARG 2002). Research on a range of related issues, such as task- and ego-involvement, motivation to learn, self-efficacy and self-theories has all shown evidence of their strong effects on learning achievement (see e.g. Butler 1988, Dweck 2000). Assessment practices which give frequent feedback to students with emphasis on scores, grades and rank-order, within a competitive culture, actually damage the motivation and achievement of many learners, particularly as they are frequently accompanied with adoption of an 'IQ culture' with its belief that one's innate and fixed intelligence cannot be altered by effort. Dweck has shown, for example that such a culture can both depress the low achievers, who believe there is no point in trying, and can damage high-attainers, who come to avoid new challenges lest they reveal that they are not as smart as they thought they were. Such findings indicate that assessment feedback should, wherever possible, be formative – giving each individual guidance about how he or she can improve and thereby fostering the belief that effort can enhance one's power to learn.

Communities

To varying degrees, learning happens through, and in association with, social interactions. Thus the learning of students develops in several communities – notably the family, the peer group, and the classroom. The idea that the nature of the classroom as a learning community is an important determinant of school learning is now well established. For example, development of peer learning and peer assessment have been shown to lead to significant improvements in learning (Black & Wiliam, 1998); it follows that teachers ought to be skilled at developing peer group work in ways that are appropriate to the subject discipline. This has implications for practices of formative assessment. The elicitation of evidence, and the formulation of feedback in response to this evidence, have to be conducted mainly at the level of the whole classroom or with small groups within a classroom – a teacher cannot have the time to conduct one-on-one interactions with all of a class. Such practice calls for development of tools, by teachers and for students themselves, to enrich students’ capacities for co-operating in formative peer assessment (Black et al. 2003).

Two questions follow for summative assessment. One is whether the day-today contributions in peer learning could, and/or should, be recorded and be part of an aggregation of evidence over time to inform teachers’ summative assessments. The other is whether there should be formal summatively -oriented exercises in which contributions to group learning are assessed in controlled contexts in order to ensure comparability across classrooms and teachers.

Teachers learn from one another, in informal groups within schools, within the formal structures of their school, in INSET training groups, and in professional associations. Many studies of learning in the contexts of trade, business and the other professions have shown that such communities of learning are formed which can have powerful effects on professional practice. Such communities need to be studied seriously where strategies for improvement of professional practice in assessment are under consideration.

Public and Policy Issues

Assessment methods provide tools that can be used in a variety of ways. The choice and deployment of these tools, and the interpretation and use of their results are subject to a range of educational, public and political influences. The variety and complexity of these influences may be listing some of them as follows :

- beliefs about what constitutes learning;
- beliefs in the reliability and validity of the results of various tools;
- trust in the objectivity of formal testing;
- a preference for and trust in numerical data, with bias towards a single number;
- trust in the judgements and integrity of one’s children’s teachers;

- trust in the judgments and integrity of the teaching profession as a whole;
- belief in the value of competition between students;
- belief in the value of competition between schools – the market model of education;
- belief that test results are a meaningful indicator of school effectiveness;
- fear of national economic decline and belief that education is crucial to improvement;
- belief that the key to schools' effectiveness is strong top-down management.

All of the above are arenas of contention, and each may reflect beliefs that are neither based on evidence nor susceptible to change by arguments from evidence. The various elements interplay in many and complex patterns which are embedded in a national culture as a whole. Safe generalizations are hard to come by, and understanding might only be enriched by case studies of individual countries.

Thus, in the case of France, the *baccalaureat* has deep historical roots in the Napoleonic constitution (Broadfoot 1996), and thereby enjoys strong public confidence with accompanying resistance to radical change. The avenues to change have been evolutionary, and both regionalization and diversification have changed its character in response to perceived economic and employment needs. Its shortcomings in terms of reliability and validity are not explored nor even questioned. However, the belief that improvement in learning depends on the commitment of and respect for the teaching profession has informed national policies which combine sampling surveys to inform judgements with provision of tools to help teachers to improve their assessments.

The case of New Zealand exhibits interesting differences and similarities. There is no tradition except of the country's past colonial dependence, which produced a tendency to follow England in its curriculum and assessment reforms of the 1990s. However, the resolution of the struggles between right and left wing policies has played out differently, and distrust of teachers and of academies has been a less powerful influence. So blanket testing of all at several ages has only hovered on the edge of the agenda, whilst support of teachers and the uses of surveys have been well-developed. For end-of-school summative assessment, the country has been more ready than the UK to radically reform its system to provide one that reflects perceived employment and economic needs.

Germany resembles France in having a tradition, albeit more recent, of a national test, the *Abitur*, the status and influence of which are under strain because of the pressures of mass education. The trust given to teachers varies between different regions. At the same time, the practices of tracking, of repeating the year, and of strong differentiation between different types of secondary schools, reflect a distinctive set of educational and social beliefs, and put unique pressures on the summative judgments of teachers.

England is different again. Here, the revolution put into effect in the 1990s reflected deep distrust of teachers combined with fear that education was a cause of economic decline.

The trust in formal tests as an engine of change has led to pupils in England stealing from the USA the dubious distinction of being the most frequently tested in the world.

However, the test results have shown the typical pattern of initial improvement followed by saturation, so frustrating politicians' promises of continued improvement. Signs of change in national policy are slowly emerging, with trials to replace some national testing by teachers' assessments, and initiatives to promote formative assessment.

Social trends can also lead to new demands on any testing system. Emphasis on the need for, and advantages of, mass higher education may reflect, and help create, demands that cannot be satisfied – so selection for places has to be made more difficult. In France, those seeking entry to the most prestigious university institutions, the *grandes écoles*, have to spend two further years after the baccalauréat to prepare for further selection tests – the majority fail and then enroll in institutions that require no more than a baccalauréat. In Germany, the *Abitur* used to guarantee automatic entry to university courses, but pressure on places in the popular courses has led universities to introduce their own supplementary tests. In England, the rise in both numbers attempting and the proportions succeeding in the examinations which are used for university selection (the A-level), is leading to pressure for a finer grain of reporting for the top grades, whilst some prestigious universities also plan to use their own 'aptitude' tests to supplement the public test results.

All of this illustrates several lessons. One is that in each country assessment practices have impacts on teaching and learning which may be strongly amplified or attenuated by the national context. Indeed, the overall impact of particular assessment practices and initiatives is determined at least as much by culture and politics as it is by educational evidence and values. A further lesson is that it may be idle to draw up maps for the ideal assessment policy for a country, even although the principles and the evidence to support such an ideal might be clearly agreed within the 'expert' community. The way forward might rather lie in those arguments and initiatives which are least offensive to existing assumptions and beliefs, and which will nevertheless serve to catalyze a shift in them while at the same time improving some aspects of present practice.

9. References

Airasian, P.W. (1991) *Classroom Assessment*, New York, McGraw Hill. Chapters 6 and 7

American Educational Research Association (AERA) (1999), with APA (American Psychological Association) and NCME (National Council on Measurement in Education) *Standards for educational and psychological testing*. Washington DC, AERA.

Angoff, W. H. (Ed.) (1971). *The College Board admissions testing program: a technical report on research and development activities relating to the Scholastic Aptitude Test and achievement tests* (2 ed.). New York, NY: College Entrance Examination Board.

ARG (2002) *Testing, Motivation and Learning* University of Cambridge School of Education : Assessment Reform Group

- Black, P. & Wiliam, D (1998) Assessment and Classroom Learning. *Assessment in Education*, **5**(1), 7-71.
- Black, P; Harrison, C.: & Hodgen, J. (2004). *Teachers' difficulties in implementing formative assessment*. Report to the Department for Education and Skills KS3 Initiative. London, UK: King's College London Department of Education and Professional Studies.
- Black, P.; Harrison, C.; Lee, C.; Marshall, B. & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Buckingham, UK: Open University.
- Black, P., Harrison, C., Osborne, J. and Duschl, R. (2004) Assessment of Science Learning 14-19: a report prepared for the Royal Society. London: The Royal Society
- Black, P. J. & Wiliam, D. (2004). The formative purpose: assessment must first promote learning. In M. Wilson (Ed.) *Towards coherence between classroom assessment and accountability: 103rd Yearbook of the National Society for the Study of Education (part 2)* (pp. 20-50). Chicago, IL: University of Chicago Press.
- Bransford, J.A., A.Brown, and R.Cocking. (1999). *How People Learn; Brain, Mind, Experience and School*. Washington D.C.: National Academy Press. Available on <http://www.nas.edu>.
- Braun, H. (2004). Personal communication, April.
- Britton, E. D. & Raizen, S. A. (Eds.). (1996). *Examining the examinations: an international comparison of science and mathematics examinations for college-bound students*. Boston, MA: Kluwer Academic Publishers.
- Broadfoot, P. (1986) *Profiles and Records of Achievement: a review of issues and practice*, London, Holt Saunders. See also pp.191-4 in Broadfoot, P. (1996) *Education, Assessment and Society*, Buckingham, Open University Press.
- Broadfoot, P.M. (1996) *Education, Assessment and Society; A Sociological Analysis*, Buckingham: Open University Press.
- Brown, M. L. (Ed.) (1992). *Graded Assessment in Mathematics: teacher's guide*. Walton-on-Thames, UK: Nelson.
- Brown, M. L.; Blondel, E.; Si mon, S. A. & Black, P. J. (1995). Progression in measuring. *Research Papers in Education*, **10**(2), 143-170.
- Brown, M. (1998). Formative assessment for learning: general issues illustrated by examples from England. In P. Black & A.Michel (eds.) *Learning from Pupil Assessment : International comparisons. Centre for the Study of Evaluation Monograph Series No.12*. Los Angeles CA : University of California Los Angeles.
- Butler, R. (1988). Enhancing and undermining intrinsic motivation; the effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of*

Educational Psychology, **58**, 1-14.

Cresswell, M. J. (1994). Aggregation and awarding methods for national curriculum assessments in England and Wales: a comparison of approaches proposed for key stages 3 and 4. *Assessment in Education: Principles Policy and Practice*, **1**(1), 45-61.

Darling-Hammond, L., Ancess, J. and Falk, B. (1995) *Authentic Assessment in Action Studies of Schools and Students at Work*, New York, Teachers' College Press. A study based upon detailed case-studies of innovations in teachers' assessments in five schools in or near New York. See particularly chapters 1 and 7.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. London: Taylor and Francis.

Educational Testing Service Cooperative Test Division (1957). *Cooperative Sequential Tests of Educational Progress: technical report*. Princeton, NJ: Educational Testing Service.

Eckstein, M. A. & Noah, H. J. (1993). *Secondary school examinations*. New Haven, CT: Yale University Press.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, **43**, 95-103.

Friedman, L. (1989). Mathematics and the gender gap: a meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, **59**(2), 185-213.

Gardner, J. & Cowan, P. (2000). *Testing the test: a study of the reliability and validity of the Northern Ireland transfer procedure test in enabling the selection of pupils for grammar school places*. Belfast, UK: Queen's University of Belfast Graduate School of Education.

Gauld, C.F. (1980) Subject oriented test construction, *Research in Science Education*, **10**, pp.77-82.

Harlen, W. (2004) A systematic review of the evidence of the reliability and validity of assessment by teachers for summative purposes. Research review produced for the EPPI Centre Institute of Education, London. In press.

Hyde, J. S.; Fennema, E. & Lamon, S. J. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychological Bulletin*, **107**, 139-155.

Johnson, S. (1988). *National assessment: the APU science approach*. London, UK: Her Majesty's Stationery Office.

- Kingsbury, G. G.; Olson, A.; Cronin, J.; Hauser, C. & Houser, R. (2003). *The state of state standards: research investigating proficiency levels in fourteen states*. Portland, OR: North West Evaluation Association.
- Koretz, D., Stecher, B. Klein, S. and McCaffrey, D. (1994). *The Vermont Portfolio Assessment Program; Findings and Implications*, Los Angeles, University of California - CRESST and Rand Institute.
- Lave, J. & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.
- Lemann, N. (1999). *The big test: the secret history of the American meritocracy*. New York, NY: Farrar, Straus & Giroux.
- Linn, M. C. (1992). Gender differences in educational achievement. In Educational Testing Service (Ed.) *Sex equity in educational opportunity, achievement, and testing: proceedings of a 1991 ETS Invitational Conference* (pp. 11-50). Princeton, NJ: Educational Testing Service.
- Linn, R. L. & Baker, E. L. (1996). Can performance-based student assessment be psychometrically sound? In J. B. Baron & D. P. Wolf (Eds.), *Performance-based assessment—challenges and possibilities: 95th yearbook of the National Society for the Study of Education part 1* (pp. 84-103). Chicago, IL: National Society for the Study of Education.
- Linn, R.L. (2000) . Assessments and Accountability. *Educational Researcher*. **29**,(2), 4-16.
- Martin, M. O.; Mullis, I. V. S.; Gonzalez, E. J.; Gregory, K. D.; Smith, T. A.; Chrostowski, S. J.; Garden, R. A. & O'Connor, K. M. (2000). *TIMSS 1999 international science report: findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: Boston College Lynch School of Education.
- Maxwell, G.S. (2004) *Progressive assessment for learning and certification: Some lessons from school-based assessment in Queensland*. Paper presented at the 3rd conference of the Association of Commonwealth Examination and Assessment Boards, March 2004, Nadi, Fiji.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.) *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education/Macmillan.
- Mullis, I. V. S.; Martin, M. O.; Gonzalez, E. J.; Gregory, K. D.; Garden, R. A.; O'Connor, K. M.; Chrostowski, S. J. & Smith, T. A. (2000). *TIMSS 1999 international mathematics report: findings from IEA's Repeat of the Third International Mathematics and Science Study at the eighth grade*. Chestnut Hill, MA: Boston College Lynch School of Education.

- Newmann, F.M., Bryk, A.S. & Nagaoka, J.K. (2001) *Authentic Intellectual work and Standardized tests: Conflict or coexistence ?* Chicago: Consortium on Chicago School Research.
- Pellegrino, J.W., Chudowsky, N. & Glaser, R. (eds.) (1999) *Knowing what students know: The science and design of educational assessments*. Washington, D.C.: National Academy Press. Available on <http://www.nas.edu>.
- Pirie, S. E. B. (1987). *Nurses and mathematics: deficiencies in basic mathematical skills among nurses – development and evaluation of methods of detection and treatment*. London, UK: Royal College of Nursing (Scutari Press).
- Quetelet, L. A. J. (1835). *Sur l'homme et le développement de ses facultés, essai d'une physique sociale [On man, and the development of his faculties, an essay on social physics]*. London, UK: Bossange & Co.
- Rogosa, D. (1999, July). *Accuracy of individual scores expressed in percentile ranks: classical test theory calculations*. Report prepared for US Department of Education Office of Educational Research and Improvement. Stanford, CA: Stanford University.
- Schulz, E. M. & Nicewander, W. A. (1997). Grade equivalent and IRT representations of growth. *Journal of Educational Measurement*, **34**(4), 315-331.
- Shayer, M. & Adey, P. S. (1981). *Towards a science of science teaching: cognitive development and curriculum demand*. London, UK: Heinemann Educational Books.
- Stecher, B. (1998) The Local Benefits and Burdens of Large-scale Portfolio Assessment, *Assessment in Education*, **5**, 335-352.
- Stobart et al. (2004) Evaluation report on the trials of the formative assessment programme of the Scottish Education Department. London: Institute of Education
- Tamir, P. (1990). Justifying the selection of answers in multiple choice items. *International Journal of Science Education*, **12**(5), 563-573.
- Towns, M. H., & Robinson, W. R. (1993). Student Use of Test-Wiseness Strategies in Solving Multiple-Choice Chemistry Examinations. *Journal of Research in Science Teaching*, **30**(7), 709 - 722
- William, D. (1992). Special needs and the distribution of attainment in the national curriculum. *British Journal of Educational Psychology*, **62**, 397-403.
- William, D. (1993) *Technical issues in the development and implementation of a system of criterion-referenced age-independent levels of attainment in the National Curriculum of England and Wales*. Unpublished King's College University of London PhD thesis.

William, D. & Black, P. (1996) Meanings and consequences : a basis for distinguishing formative and summative functions of assessment *British Educational Research Journal* **22** (5) 537-548

Williamson, G. L.; Appelbaum, M. & Epanchin, A. (1991). Longitudinal analysis of academic achievement. *Journal of Educational Measurement*, **28**(1), 61-76.

Wood, R. (1991) *Assessment and Testing* , Cambridge, Cambridge University Press. Chapters 1 to 5, 9,10,13 and 15 to 19.

Yen, W. M. (1986). The choice of scale for educational measurement: an IRT perspective. *Journal of Educational Measurement*, **23**, 299-325.