The Curriculum Journal
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rcjo20

Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices

Paul Black & Dylan Wiliam
[a] King's College London
[b] ETS , Rosedale Road (ms 04-R), Princeton, NJ, 08541, USA E-mail:
Published online: 17 Feb 2007.

To cite this article: Paul Black & Dylan Wiliam (2005) Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices, The Curriculum Journal, 16:2, 249-261, DOI: 10.1080/09585170500136218

To link to this article: http://dx.doi.org/10.1080/09585170500136218

The Curriculum Journal, Vol. 16, No. 2, June 2005, pp. 249–261

Routledge
Taylor & Francis Group

Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices

Paul Black and Dylan Wiliam*
King's College London

This article outlines the main assessment traditions in four countries – England, France, Germany and the United States – in order to explore the prospects for the integration of summative and formative functions of assessment during compulsory schooling. In England, teachers' judgments do feed into national assessments, at 7, 11, 14 and 16, but concerns for reliability and accountability mean that such judgments are made in a way that has little impact on learning. In France, teachers have no involvement in the formal assessment of students, and, possibly as a result, have been free to concentrate on the use of assessment to serve learning. In Germany, faith in the education system has been considerably undermined by recent unfavourable international comparisons, although faith in the ability of tests both to measure learning accurately and to allocate students to different educational pathways appears to be unshaken. In the United States, multiple demands for accountability at different levels of the system have resulted in multiple assessment systems, but these tend to be focused on measuring the amount of learning that has taken place, providing little insight into how it might be improved. It is concluded that the effective integration of formative and summative functions of assessment will need to take different forms in different countries, and is likely to be extremely difficult.

Keywords: *accountability, comparative education, formative assessment, summative assessment*

Introduction

Public schooling takes very different forms in different countries. In most developed countries, attendance is compulsory from the age of 5 or 6 to the age of 15 or 16, although many systems allow alternatives such as home schooling. In addition, most students attend some form of nursery or pre-school, and most students continue in education beyond the end of compulsory school. In most countries private schools

*Corresponding author. ETS, Rosedale Road (ms 04-R), Princeton, NJ 08541, USA.
Email: dylanwiliam@mac.com

ISSN 0958-5176 (print)/ISSN 1469-3704 (online)/05/020249–13

DOI: 10.1080/09585170500136218

operate entirely outside the state school system, but in others (Australia, for example) some private schools are supported with public money. In some countries (England, for example) parochial schools are part of the state system, while in others (such as the United States), they are outside it.

Attitudes towards selective schooling differ markedly. In most industrialized countries, the vast majority of students attend schools that are open to all; in Germany, at the end of four years of primary schooling, however, students are allocated to different types of schools on the basis of their academic achievement. In England some local education authorities also retain selection at the end of primary school, although the proportion of students attending selective secondary schools ranges from 4 per cent in some districts to over 30 per cent in others. In Japan, the most prestigious schools also select on ability but for the others there is little, if any, competition.

In England, France, Japan, New Zealand and Sweden there is a national curriculum while in Australia and Germany the curriculum is the responsibility of regional bodies. In the United States, school curricula are nominally the responsibility of approximately 17,000 directly elected local boards of education, but 49 of the 50 states (the exception is Iowa) have laid down state curriculum standards to which local curricula are required to conform.

Curricula invariably include reading and writing the mother tongue, mathematics, science and the humanities, although there are differences in how these are organized. For example, the earth sciences are treated as part of geography in England but as part of science in the United States. Chemistry and physics are separate subjects in upper secondary schools in England, but are taught together in France. Moral education is an explicit subject in some countries (e.g. Japan) but subsumed in other subjects in other countries. Religious education is compulsory in state schools in some countries (e.g. England) and illegal in others (e.g. the United States). In some countries (e.g. England, Germany and the United States) the upper secondary school curriculum allows students considerable choice in terms of what to study, while in others the choice is between pathways, within which there is little choice (e.g. France, Sweden).

Policies regarding textbooks also vary. At one extreme, in Japan, textbooks are required by law, and textbooks must be approved by the government. In Germany, each of the 16 *Länder* lays down regulations for the adoption, introduction and use of textbooks. In the United States some states (e.g. California, Florida, Texas) must approve textbooks before they can be used, while in others, textbook adoption is a matter for individual school districts. In England, the decision about which textbooks to use is a matter for the school, and sometimes even the individual teacher.

Given these differences in the fundamental characteristics of education systems, it is hardly surprising that there are also significant and deep-rooted differences in the assessment systems of different countries. One tradition in the field of comparative education is to try to draw out the common trends in different countries, thus relegating the local differences as being of secondary importance. The aim of this article is slightly different. It is to try to bring these differences into the foreground in

order to attempt to understand the assessment traditions in each country within their cultural and political contexts, and then to attempt to reflect on the prospects for the development of effective formative assessment. To do so, this article focuses on four countries, England, France, Germany and the United States, and presents a description of the key features of their assessment systems together with a discussion of how the assessment systems are influenced by their cultural and social contexts.

England

When the government announced its intention to introduce a national curriculum for England (and Wales) in 1987, it was made clear that the assessment of students' achievement at the ages of 7, 11 and 14 would be through a combination of teacher judgements and externally set tests. Currently, all national curriculum subjects are assessed by teacher judgement at the ages of 7, 11 and 14, and in addition there are formal tests for English and mathematics at ages 7, 11 and 14, and for science at 11 and 14.

This attention to the role of teachers' judgements as part of a summative assessment has obscured the role that assessments can play in supporting learning at both the practical and policy levels. In terms of practice, the lack of central guidance about what would constitute adequate records of student achievement resulted in a myriad of complex 'home-grown' record-keeping systems that were ill suited to supporting learning but provided comprehensive evidence of students' achievements. At the policy level, the failure to appreciate that the 'teacher assessment' mandated as part of the statutory assessment arrangements was entirely summative in nature led to a situation in which 'formative assessment' was discussed only twice by the government agency responsible for the national curriculum and its assessment during the first seven years of implementation (Daugherty, 1995). Even recent pronouncements about the importance of assessment for learning have tended to reinforce the collection, rather than the use, of data.

Teachers' judgements also feature in the national assessments for 16- and 18-year-old pupils. There are two main qualifications systems for students in this age group. The traditional route, and the one taken by most students aspiring to go on to higher education, has been to take the General Certificate of Secondary Education or GCSE at age 16 and the General Certificate of Education Advanced Level (usually abbreviated to 'A-level') at age 18.

The GCSE typically takes the form of a combination of timed written examination papers (one or two papers for each subject, lasting up to two-and-a-half hours each) and assessment by teachers of work completed by the student in school and at home. The weight of the school-based component ranges from 20 per cent (in mathematics) up to 50 per cent (in technology), and the results are graded on a nine-point scale: U, G, F, E, D, C, B, A, A* (highest).

For the final two years of secondary education, students following the academic route specialize in a small number of subjects (typically four or five) and, after a year's study, most specialize further by concentrating on just three subjects for the final year

of schooling, at the end of which they take the A-level examinations. Alongside these academic qualifications, and within the same integrated qualifications framework, is a range of vocational qualifications. When originally developed in the 1990s, the vocational qualifications made much greater use of teachers' judgements in determining final outcomes, and many were designed specifically so that the assessment evidence generated for summative purposes could also be used formatively. However, as these vocationally oriented programmes were integrated into a comprehensive qualifications framework, assessment practices moved much more towards a model in which test and examination scores are combined with teachers' assessments of formal 'set pieces' undertaken during the programme.

Thus, in the 5 – 14 general curriculum, and in both academic and vocational routes for 14 – 19 year olds, the involvement of teachers in the assessment process originally held out the promise that the assessments undertaken by teachers as part of their normal teaching could serve both formative and summative functions. However, by insisting that these teacher assessments should be judged according to the definitions of reliability that had been developed for traditional tests, combined with a profound lack of trust in teachers, the potential for teacher assessments to support learning has been substantially eroded. Some teachers have been able to develop assessment methods that integrate both formative and summative functions effectively, but, even here, the emphasis has been on maximizing test scores rather than meeting students' learning needs, so that, in a parody of Gresham's law, summative has driven out formative. As one teacher put it: 'It is a bit depressing, isn't it?' (Black & Wiliam, 2004).

France

In France, there is a national curriculum for all levels of compulsory education that also extends to pre-school and post-compulsory education. Schools do have a limited degree of flexibility in the allocation of time between different subjects, and teachers are expected to use their own creativity to bring out the best in every student. At the age of 15, students can take the *brevet des collèges*, or for those in vocational programmes, the *certificat d'aptitude professionelle* (CAP) and the *brevet d'études professionelles* (BEP). Originally intended as a school-leaving examination, a pass in the *brevet* was needed for entry into the *lycée* but is no longer required. However, the *brevet* is now seen as good practice for later examinations, and although voluntary, is taken by 99 per cent of students in the last year of the *collège*, and 75 per cent achieve a pass.

The main requirement for entry to higher education is a pass in the *baccalauréat*, taken at the end of the third year of the *lycée*. This has three variants: *général* (academic), *technologique* (technical) and *professionel* (vocational). Although originally intended only for the highest achieving students, the proportion of students taking the *Bac* has risen over the last thirty years from 30 per cent to 90 per cent, and about 75 per cent obtain some sort of pass. With the exception of foreign languages, all subjects are assessed almost exclusively through a series of examination papers, each lasting

from three to three-and-a-half hours, and taken over a period of four days. For most universities, entry is based solely on the performance on the *Bac*, but for admission to the *institut universitaire de technologie* (polytechnic), school reports are also taken into account. For students wishing to enter the *grandes écoles* – prestigious engineering schools – students take highly competitive entrance examinations. Students prepare for these examinations by attending *classes preparatoires* at institutions that are administered within the secondary school system. If students fail to gain admission to one of the *grandes écoles*, some, but not all, universities will take the scores on the entrance examination in lieu of the *Bac*.

Unlike England, the United States and, more recently, Sweden, in France the results achieved by students on external high-stakes tests are not seen as a good way of monitoring standards of achievement in schools (although newspapers do publish 'league tables' based on schools' results in the *baccalauréat* examinations). Instead, the Ministry of Education, through its Office for Assessment and Forecasting, monitors all aspects of educational provision, including facilities and resources, classroom practices, students' achievements and school effectiveness, through focused surveys. For example, the achievement of students at the end of study in *collèges* is evaluated by use of sample surveys. As well as measuring achievement on each of the school subjects, these surveys also assess non-cognitive aspects, including attitudes and values. In addition, there are cohort studies in which samples of students are followed through several years of their schooling, so that long-term trends can be monitored.

As well as these evaluative studies, the Ministry of Education also believed that it was necessary to provide individual teachers with data on their own students:

> This is based on the belief that such stakeholders, among whom teachers are prominent, will improve their professional practices only if they are shown, as in a mirror, the consequences of their actions. (Bonnet, 1997, p. 299)

Accordingly, a system was introduced of testing all students in alternate years at the ages of 8 and 11, and every year in all subjects for students at age 16. The formative purpose of these assessments was emphasized by having the tests set at the *beginning* of the school year, so that they inform each teacher about their new class. Accordingly, these tests are seen as an aid to teaching, rather than a judgement of teachers.

Although the primary purpose of these tests was to inform teachers about their students, a sample of the data is analysed by the Ministry of Education in order to inform pre-service and in-service training programmes. In 1989, for example, it was discovered that the performance of primary school students in geometry appeared to be weaker than that in arithmetic, even, though both aspects of mathematics were meant to be given equal emphasis. The publicity given to this finding was enough to prompt teachers to place more emphasis on geometry, which subsequently led to improvement in geometry scores.

Finally, there is another layer of evaluation provided by the system of school inspectors, which is mainly devoted to evaluating the performance of individual teachers rather than to schools as such.

What is perhaps most remarkable about the French system is the plethora of different assessment and evaluation systems, each designed to serve a very limited range of functions. The assessment of students' attainment is achieved through examinations such as the *brevet* and the *Bac* but the evaluation of the system is achieved through other, purpose-built, assessments. The quality of the performance of individual teachers is assessed through the inspection system but schools are expected, in addition, to analyse their own performance against national and regional norms, including the use of 'value-added' measures to examine the progress of individual students as they move through the school. But the results on the summative tests are not seen to provide evidence about the performance of the schools, for this would 'most certainly antagonize the teaching profession – it would be seen as a way to evaluate the teachers themselves – and alter beyond recognition the formative nature of mass assessment' (Bonnet, 1997, p. 303).

While it would be going too far to say that the use of traditional, timed examinations is regarded as unproblematic in France, there certainly seems to be a strong and widespread belief that such assessments are fair and valid ways of assessing students. More importantly, the fact that the teacher has no role in assessing the student summatively leaves the teacher free to concentrate on learning, and it is therefore probably not a coincidence that the role of assessment in the support of learning has been such a strong aspect of French teaching for many years. While the term '*evaluation*' in French clearly includes what would be termed 'formative assessment' and 'summative assessment', the lack of a role for teachers in the latter appears to have resulted in the former being fully integrated into pedagogy. Indeed, as the French research literature on this topic makes clear, the term '*évaluation formative*' includes all the ways teachers might elicit from students evidence of their understanding, including discussions with and observations of students (Allal & Lopez, 2005).

Germany

The responsibility for education policy in Germany rests with each of the 16 *Länder*, but through the Conference of Education Ministers (KMK) agreement has been reached on a wide range of issues in education, including the curriculum, grading systems used in school, and mutual acceptance of each other's qualifications.

In grades 1 and 2, parents are given report cards at the end of each year which detail their child's work habits, special skills and weaknesses, behaviour, attitude to learning, and their participation in class, but no grades are given. No student is required to repeat these grades unless this is requested by the parents (for example, in the case of a student who has an extended absence due to illness).

In the higher grades, the results of the formal tests and examinations are reported to parents, who must sign to show that they have seen them. Any student getting one of the two lowest grades (5 or 6) in two subjects for the year can be required to repeat the year. Any student who is held back twice is evaluated to see whether they would benefit from attending a special school. Officially, 4 per cent of students in a cohort

are held back, with the cumulative effect being that by the age of 14 as many of 20 per cent of students are being taught in classes where the majority of students are at least a year younger (Beaton *et al.*, 1996a, 1996b).

During the final year of elementary school (grade 4), on the basis of the student's performance, teachers recommend the appropriate form of lower secondary school (in one *Land* there is a central examination). If parents are not happy with the recommended allocation they can appeal, but the burden of proof rests on them to show that the student has the necessary ability to prosper in the chosen school (for example, through additional tests and examinations).

The teachers' recommendations are based largely on the scores that students achieve on a series of tests and examinations administered during the year, the format, length and frequency of which are determined by the Ministry of Education. For example, in one *Land*, the Ministry of Education specifies that only German and mathematics may be assessed formally, on the following format, with each subject being assessed according to the following schedule (Nerison-Low, 1999):

> 2nd grade: 4 tests, no more than 15 minutes each;
>
> 3rd grade: 6 examinations, 3 no more than 30 minutes and 3 no more than 15 minutes;
>
> 4th grade: 6 examinations, 4 no more than 30 minutes and 2 no more than 45 minutes.

Teachers typically collaborate to ensure that the tests and examinations that they set make comparable demands on students, to agree how the tests will be scored, and how the marks on the test will be converted to the six-point scale used throughout the German education system.

However, the German educational system is currently in a great deal of flux. German students displayed strong performances in the first and second international mathematics and science studies (Brown, 1996), but in the third study, which controlled more effectively for the effects of grade retention, German students were shown to be considerably weaker than in the previous studies. Moreover, the correlation of achievement with social class was among the strongest of all the industrialized countries. As Thürmann (2004) notes, 'The media interpreted these results as nothing short of a national catastrophe', and a variety of measures have been put in place to improve the performance of the school system.

The general trend of the reforms is reminiscent of those introduced in England and Wales in the 1988 Education Reform Act. Schools are to become more accountable through regular monitoring of performance, but are also to be given greater autonomy to find their own ways to improve.

The measures to increase accountability have included the adoption, in 2003, of common curricular standards for the core subjects across all 16 *Länder*, and many *Länder* have introduced centralized assessment systems to ensure that standards are being applied consistently, and that schools are using data to monitor students' learning. For example, in North Rhine–Westphalia recent reforms have introduced:

new curricular standards in German, mathematics and foreign languages;

school-based 'parallel assessment' for a selection of subjects in grades 3, 7, 10, 11/12;

state-wide mandatory assessment for core subjects in grade 4 and at the beginning of grade 9;

centrally set examinations at the end of lower secondary school;

centrally set examinations at the end of upper secondary education;

independent school inspection. (Thürman, 2004)

At the same time, schools are expected to put in place systems of self-evaluation in order to find their own ways to improve.

Given these radical changes, what is perhaps most surprising to an outside observer is that, at the moment, faith in the core features of German education seems relatively unshaken. The research on the effects of selection seem to indicate that the use of selective systems increases the variance in outcomes, while depressing the mean (Bursten, 1992), and yet there are few calls for an end to the tripartite system of secondary schooling. Similarly, while the research on grade retention is controversial (see Alexander *et al.*, 1995), the use of grade retention adds substantially to the cost of educational provision, and other ways of using these resources are likely to be more effective. Finally, and most importantly, there appears to be substantial faith in the usefulness of regular testing both to motivate students and to provide useful information to teachers, despite the lack of evidence to support this (Köller, 2005).

Nevertheless, it is important to note that education in Germany is undergoing its most radical upheaval for at least fifty years. As Thürmann notes, 'The acceleration of change and the almost complete change of perspective from input-orientation to output-control and evaluation is breath-taking' (2004, p. 14). Even such hallowed features as selection and grade retention may come under scrutiny as schools come under greater and greater pressure to improve the achievements of their students, although the faith in regular summative testing as the best way to achieve this would appear to remain very strong.

United States

The most important feature of the education system in the United States is that there isn't one. It is much more productive to think of the United States as one might think of Europe, so that there are 51 systems (50 states plus the District of Columbia), and in many respects there are 17,000, since each school district has considerable autonomy in determining school organization, curricula, teachers' pay, and the day-to-day operation of its schools. The largest school district, New York City Public Schools, has over 1 million students – more than many countries – while the Lake Alice School District in Nebraska has just 1 school, 6 teachers and 77 students.

Like France and Germany, the United States operates a grade-based curriculum, but grade-retention, although increasing, is comparatively rare in grades K–8, so that while the curriculum is designed as a grade-based system, it is not operated as such.

This, combined with the widespread use of mixed-ability classes in these grades, may well be why the same material tends to be repeated in the curricula of more different grades in the US than in other countries (Schmidt *et al.*, 1997) and, in mathematics at least, more than half of lesson time is spent on reviewing previously taught material (Hiebert *et al.*, 2003).

Beginning in the third or fourth grade (and continuing through to postgraduate level!), almost all formally assessed student work is assessed on the same literal grade scale: A, B, C, D, F (fail), typically corresponding to percentage scores of 90–100, 80 – 89, 70 – 79, 60 – 69 and 0 – 60 respectively. Grades are cumulated by converting them back to numbers (A = 4, B = 3, C = 2, D = 1, F = 0) and calculating the 'grade-point average' over the year. However, unlike scores or grades given in most European countries, the grade is usually not a pure measure of attainment, but will include how much effort the student put into the assignment, attendance, and sometimes even behaviour in class. Paul Dressel's definition of a grade was 'an inadequate report of an inaccurate judgement by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material' (Chickering, 1983), and while this may be a bit unfair, there can be little doubt that the meaning of a grade varies substantially from school to school, and even from teacher to teacher.

Nevertheless, great importance is attached to grades, both as an indication of the progress a student makes at school, and as 'currency' for applications to higher education institutions, and teachers frequently come under pressure to modify their grades if they are felt by parents to be too low. While only a few districts have gone as far as establishing 'grade courts' where parents can take grievances about grades, and where teachers have to defend the grades they have awarded, teachers do feel under considerable pressure to be able to defend the grades they have awarded and grades are often just based on the outcomes of objective tests.

The other feature of education in the United States that militates against the introduction of effective formative assessment is the response to increases in testing for accountability purposes. Most of the money needed to operate schools is provided by taxes on residential and commercial property within the school district, and most districts are governed by boards elected by local taxpayers. Over the last forty years or so, however, state and federal sources have become greater and greater net contributors (Corbett & Wilson, 1991, p. 25), leading to demands that school districts become accountable beyond the local community. As noted above, all states but Iowa have state-wide educational standards, and most have also implemented some form of state-wide testing programme.

Although some states implemented systems of assessment that threatened sanctions for individual students for poor performance (e.g. requiring that students pass minimum-competency tests before they can be awarded high school diplomas), most state systems were, and remain, low stakes for students but high stakes for the schools. This principle is continued in the No Child Left Behind Act (NCLB). Under NCLB, each state must propose a series of staged targets for achieving the overall goal of all students in grades 3–8 being proficient in reading and mathematics by 2014.

Each school is judged to be making 'adequate yearly progress' (AYP) towards this goal if the proportion of students being judged as 'proficient' on annual state-produced tests exceeds the target percentage for the state for that year. Furthermore, the AYP requirements apply not only to the totality of students in a grade but also to specific subgroups of students (e.g. ethnic minority groups), so that it is not possible for good performance by some student subgroups to offset poor performance in others.

Failure to make AYP has severe consequences for schools and, as a result, many schools and districts have invested both time and money in setting up systems for monitoring what the teachers are teaching and what students are learning. In order to ensure that teachers cover the curriculum, most districts have devised 'curriculum pacing guides' that specify which pages of the set texts are to be covered every week (and sometimes each day). With such rigid pacing, there are few opportunities for teachers to use information on student performance to address learning needs.

More recently, there has been a huge upsurge of interest in systems that monitor student progress through the use of regular formal tests that are designed to predict performance on the annual state tests. The idea of such regular testing is that students who are likely to fail the state test, and may therefore prevent the school from reaching its AYP target, can be identified and given additional support, and for this reason, these systems are routinely described in the USA as 'formative assessment', even though the results of the assessments rarely impact on learning and, as such, might be better described as 'early-warning summative'. The result is that while teachers in the United States would appear to have considerable freedom to devise assessment systems that integrate summative and formative uses, the obstacles to doing so are substantial, and well entrenched.

Conclusion

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 seen by 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 judgements 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 that are embedded in a national culture as a whole. Safe generalizations are hard to come by, and any deep understanding may come only from case-studies of individual countries.

Thus, in the case of France, the *baccalauréat* has deep historical roots in the Napoleonic code (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 or 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.

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 judgements 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.

In the United States, teachers have been trusted by their local communities (as the annual *Phi Delta Kappan* surveys show) but not by policy-makers at the state and federal level. The result is an uneasy compromise, with districts free to decide what their students learn, and how student performance is to be assessed while, at the same time, holding schools accountable for students' performance on material that they may not have been taught.

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 enrol 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) are leading to pressure for a finer grain of reporting for the top grades, while 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 that 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 is likely to be idle to draw up maps for the ideal assessment policy for a country, even if the principles and the evidence to support such an ideal could be clearly agreed within the 'expert' community. The way forward might, rather, lie in those arguments and initiatives that are least offensive to existing assumptions and beliefs, and which will nevertheless serve to catalyse a shift in them while at the same time improving some aspects of present practice.

As far as the integration of formative and summative functions of assessment is concerned, the analysis of the four national systems described above is something of a paradox: the better the teacher knows her or his students, through processes of formative assessment, the less likely it is that the information is used to inform judgements made about the student. In France, albeit under the banner of pedagogy rather than assessment, effective classroom assessment practices have been developed on a widespread basis, but the key decisions about a student's educational trajectory are made on the basis of examination results. In Germany, teachers are involved in making key decisions about student progress but the primary evidence base for this is provided by formal tests. Similarly, in England, teacher judgements do feed into national assessments, but the demands of teachers to be accountable for these judgements has resulted in record-keeping systems that do not serve learning well. Finally, in the United States, multiple demands for accountability at different levels of the system have resulted in multiple assessment systems, all geared to serving the summative function of assessment, so marginalizing, and denying time to, assessment that supports learning.

Thus not only is there no 'royal road' to an assessment system that effectively serves both formative and summative functions that each country could follow, but it seems likely that the idiosyncratic road that will need to be taken in each country will also be very hard going. The final irony is that it is precisely the demand for accountability which has produced unprecedented pressure to improve education systems that is likely to be the biggest impediment to achieving that improvement.

Note

Paul Black is Professor Emeritus of Science Education at King's College London. Dylan Wiliam is Director of the Learning and Teaching Research Center at ETS, Princeton, New Jersey, but writes here in a personal capacity.

References

Allal, L. & Lopez, L. M. (2005) Formative assessment of learning: a review of publications in French, in: J. Looney (Ed.) *Improving learning through formative assessment: cases, policies, research* (Paris, France, Organisation for Economic Co-operation and Development).

Alexander, K. L., Entwistle, D. R. & Dauber, S. L. (1995) *On the success of failure: a reassessment of grade retention in the primary grades* (Cambridge, Cambridge University Press).

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Kelly, D. L. & Smith, T. A. (1996a) *Mathematics achievement in the middle school years: IEA's third international mathematics and science study* (Chestnut Hill, MA, Boston College).

Beaton, A. E., Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A. & Kelly, D. L. (1996b) *Science achievement in the middle school years: IEA's third international mathematics and science study* (Chestnut Hill, MA, Boston College).

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) (Chicago, IL, University of Chicago Press), 20–50.

Bonnet, G. (1997) Country profile from France, *Assessment in Education*, 4(2), 295–306.

Broadfoot, P. M. (1996) *Education, assessment and society: a sociological analysis* (Buckingham, Open University Press).

Brown, M. L. (1996) FIMS and SIMS: the first two IEA international mathematics surveys, *Assessment in Education: Principles, Policy and Practice*, 3(2), 193–212.

Bursten, L. (Ed.) (1992) *The IEA study of mathematics III: student growth and classroom processes* (Oxford, Pergamon).

Chickering, A. W. (1983) Grades: one more tilt at the windmill, *American Association for Higher Education Bulletin*, 35(8), 10–13.

Corbett, H. D. & Wilson, B. L. (1991) *Testing, reform and rebellion* (Hillsdale, NJ, Ablex).

Daugherty, R. (1995) *National curriculum assessment: a review of policy 1987–1994* (London, Falmer).

Hiebert, J., Gallimore, R., Garnier, H., Bogard Givvin, K., Hollingsworth, H., Jacobs, J., Chui, Miu-Ying A., Wearne, D., Smith, M., Kersting, N., Manaster, A., Tseng, E., Etterbeek, W., Manaster, C., Gonzales, P. & Stigler, J. (2003) *Teaching mathematics in seven countries: results from the TIMSS 1999 video study* (Washington, DC, National Center for Education Statistics).

Köller, O. (2005) Formative assessment in classrooms: a review of the empirical German literature, in: J. Looney (Ed.) *Improving learning through formative assessment: cases, policies, research* (Paris, Organisation for Economic Co-operation and Development).

Nerison-Low, R. (1999) Individual differences and the German education system, in: M. A. Ashwill (Ed.) *The educational system in Germany: case study findings, June 1999* (Washington, DC, National Institute on Student Achievement, Curriculum, and Assessment, Office of Educational Research and Improvement, US Department of Education, National Center for Education Statistics).

Schmidt, W. H., McKnight, C. C. & Raizen, S. A. (1997) *A splintered vision: an investigation of US science and mathematics education* (Dordrecht, Netherlands, Kluwer Academic Publishers).

Thürmann, E. (2004) New curricular formats and classroom development: a change of paradigm for the German school system?, in: J. Letschert (Ed.) *The integrated person: how curriculum development relates to new competencies* (Enschede, Netherlands, CIDREE).