A Comparative Perspective on Educational Standards

ALISON WOLF

Introduction

OVER THE LAST TWO CENTURIES, the world's education systems have developed along paths which combine major, over-arching similarities with substantial and enduring differences in philosophy and organisation. This applies to the operation of 'educational standards' as much as it does to other aspects of education. Countries all share the need to both select and certify; they chase the grail of economic growth-through-education; and they treat international surveys as a sort of mini-Olympics involving national pride and the urge to beat traditional enemies. Nonetheless, there remain profound differences between countries in how educational standards operate; and these both derive from and help sustain differences in the institutions and the values of the societies concerned.

Before turning to standards themselves, let me give some more general examples of how historical trends subsume enduring differences in countries' education systems.. Across the world, inclusive, publicly funded systems of primary and secondary education are now the norm. However, there are major differences in how far they were created by public authorities from scratch, or involved the nationalisation and integration of pre-existing voluntary provision. There are differences in the degree of autonomy that component groupings (e.g. religious schools) and individual institutions enjoy. And there are also major differences in the scale and nature of independent institutions, surviving alongside publicly funded provision.

A second example of a general trend involves the way in which

Proceedings of the British Academy, **102**, 9–37.

the dominant nineteenth century pattern of education, with a tiny lycée/gymnasium/grammar school sector, teaching classics and maths, and a unified elementary school leading into the labour market at age 13 or 14 has been replaced, throughout the developed world, by universal secondary schooling. Even England is now following the new norm of full-time participation well beyond the statutory leaving age. Yet here too major and enduring differences remain; in the extent to which secondary provision is unified or selective, in the relative importance of work-based apprenticeship routes, in levels of centralised curriculum control, and in the nature and impact of formal qualifications at this level. In a recent examination of trends within the EU, colleagues and I found no evidence that secondary level education systems were converging in anything other than size and inclusiveness (Green, Leney and Wolf 1997; Green, Wolf and Leney 1999).

A similar picture applies when we look at qualifications and examinations, and so enter the domain of 'standards'. Throughout the world—and not just the developed world this time—formal assessments and qualifications are increasing in number and importance (Dore 1996, Little and Wolf 1996, Little 1996). There are increasing government anxieties over education, because of its supposedly critical contribution to economic growth and competitiveness, and increasing individual (especially parental) anxieties, because of the growing importance of formal qualifications in determining life-chances. The management and delivery of qualifications is itself a sizeable industry.

The International Association for Educational Assessment, for example, brings together school examination and assessment bodies from over 50 countries: and participants find immediate common ground because of the very limited number of ways in which this business is organised. For the majority of the world's population (including the citizens of China and India), secondary education is dominated by formal educational qualifications that involve examinations set by publicly constituted or recognised examination boards, operating with a structure of national curricula, subject committees, common papers written under 'exam conditions' on common dates, and paid anonymous examiners (hired largely from the teaching force). For most others (and most notably citizens of the United States), the key experience is of sitting standardised tests in largely multiple-choice format, which are created and (machine) marked under conditions of tight security by the permanent staff of testing agencies, relate to

general skills and traits rather than specific syllabi, and yield percentile scores related to national norms.

It would seem, therefore, that we have a bipolar global consensus on how to assess student achievement (or attainment of 'standards') for public reporting purposes. To this we have recently seen added a more ambitious effort, that of 'global' assessment. Here, it would seem, we really do have a common approach to the measurement and comparison of standards, as countries administer the same tests to all their students. By far the biggest player here is the IEA—the International Association for the Study of Educational Achievement. (See especially Goldstein 1995, Goldstein 1996.) The first IEA study in 1964 was a maths survey involving 12 countries and one age group (13 year olds): while for the recently completed TIMSS survey (Third International Maths and Science Study), three age groups were covered and between 40 and 50 countries participated in any given part. Many others were involved in early stages but were unable to provide the full samples required for formal reporting by the IEA. Moreover, such studies are multiplying. The IAEP (International Assessment of Educational Progress) treads similar ground to the IEA; IALS—the International Adult Literacy Survey—concerns itself with the 16–65 year old population; and the OECD is now preparing another in-depth survey of secondary students' achievement (the PISA study) involving the richer developed countries who make up its membership.

When a country participates in one of these studies, it is effectively signing up to a number of propositions: that there is a construct—mathematics achievement, literacy, or whatever—which has the same meaning for all the countries involved; that this can be measured in an equally valid way in all cases; and that the measurements can be reported in the form of a scale on which participants can be not merely ranked (higher/lower) but ascribed a mark or score which designates distance. That is, the implication is that in some meaningful way the difference between a score of (say) 45 and 55 is the 'same' as between one of 53 and 63. All of these are in fact highly questionable propositions (see especially Goldstein 1996b, McLean 1996). However, the relevant point here is that the growing number of countries participating in such studies would seem to imply a corresponding level of consensus on the nature of educational standards.

It is the argument of this paper that there is, in fact, far less of a clear consensus than the spread of international surveys, and the limited range of assessment and examining approaches, would imply.

It is not just a case of different conceptions of standards being associated with the three main approaches to 'public' assessment which have been described—the syllabus-related exam board, the free-standing standardised test, and the international achievement survey. Once one examines individual countries in detail, great variability becomes apparent even among countries which use the same basic approach. These differences, in turn, relate to profound differences in basic values, beliefs and objectives. The following pages demonstrate this by examining a number of exemplar countries: China, the United States, Sweden, Germany and France.

China

As most readers will know, the first system of public examinations which we would recognise clearly as such was developed by Imperial China in order to select entrants to the bureaucracy of mandarins which ran the country. Use of examinations for selecting public servants began in the Sui dynasty (606) and continued until 1905. The examinations were open to all; were set in relation to a fixed syllabus; and taken under strict 'exam conditions'—uniform for all candidates, with questions prepared secretly, and marked confidentially by experienced examiners.

China remains a culture in which public examinations remain critical to individual success, and are respected and trusted by the population. It is also a culture in which one particular examination dominates, though it is now the Entrance Examination to Higher Education rather than the examination for entry to the Imperial civil service. Though abolished temporarily during the Cultural Revolution, the EEHE has otherwise, since 1952, been by far the most important examination in P.R. China; and the general view, shared even by academics with a critical perspective on assessment, is that 'It is universally acknowledged in China that the EEHE is beneficial to the efficiency and quality of selection and plays an important role in promoting the quality of secondary education.' (Wang Gang 1995: 3).

The EEHE is set by the National Education Examination Authority in Beijing; an organisation with a network of subject committees and officers which would be immediately recognisable to anyone from a UK exam board. It is sat at the same time throughout the country; and marked by university and secondary school teachers who work together, using set marking criteria, in a seaside resort which has been taken over

for the purpose.[1] Given a school system with enrolments of 191 million pupils, this seems extraordinary, although the pyramidal nature of Chinese education means that the numbers sitting the EEHE are 'only' 2.5 million (competing for just under a million places). Nonetheless, the view of the NEEA is that any move to wholesale decentralisation would undermine faith in the examination, and create genuine dangers of corruption and political interference in the marking (personal communication).

In the last 10 to 15 years—partly because of access to Western assessment literature and debates—issues of quality and technical procedures have been debated increasingly within China; while there has also been active discussion about the nature of the examination and its backwash effects, notably about the relative emphasis on rote learning at the expense of problem-solving and critical skills. There has also been a steady refinement of the procedures for setting papers in terms of the responsibilities of subject committees, the checking of items, the pre-definition of item types, and the clarity of mark schemes.

However, the process of setting and marking papers remains fairly detached from the type of technical procedures common to American psychometrics and, to an increasing extent, English public examinations. Trialling is underway of a procedure for standardising individual subject scores before they are added up to provide the single, total score actually used for university admission. (At present, simple raw scores are used, even though different subjects tend to have markedly different means and standard deviations. See Little, Wang and Wolf, 1995.) However, there are no formal procedures in place to examine or secure standards from year to year. Since the composition of subject committees changes frequently, the main sources of stability are the well-established expectations of the student and teacher population, enshrined in the limited number of approved textbooks, and maintained through the publication of exam papers, and the expertise of committee chairmen and NEEA staff.

Ensuring strict comparability of standards from year to year has never been an issue in China, and the functions of the examination make it unlikely that it will become so. The EEHE is a selection exam: it exists in order to decide which students will be admitted into which universities.: its concerns are with 'differentiation and discrimination'

[1] There is one exception: Shanghai currently is permitted to set and administer its own version of the EEHE.

(Wang Binhua 1995: 32). According to the marks received, students may win admission into one of the 'key' universities which recruit nationally, or into a less desirable regional (provincial) university. The whole process is centrally steered, in the sense that the national ministry decides, every year, the total number of student enrolments which will be allowed from each province for both the provincial and the key universities. The score line above which students are eligible for the key universities varies among provinces, in order to ensure that some students are admitted from poorer, rural provinces, rather than simply from the Eastern seaboard; but with this proviso, the process is simply one of admitting the highest scoring candidates. A similar situation obtains in the case of secondary schools, where there is fierce competition for entry into the selective 'key' schools: and where entry is again decided totally on the basis of students' (raw) scores on locally set examinations.

There are two reasons why, in this context, year-on-year standards have not been an issue. The first is that, for the overwhelming majority of stakeholders—students, parents, politicians, university administrators—the only relevant issue is whether the examination appears to treat *a given year's entry* fairly. Given the enormous effort that is put into developing individual items, the clear expectations about question format and content which inform this process, and the tight security surrounding the whole process, this is not usually experienced as a problem: and an individual's relative ranking is then perceived to be the result of an objective decision. Because no claims are being made about substantive achievement levels, a purely selective examination of this type avoids many of the problems associated with the notion of 'standards over time'.

The other reason why these are not a major concern is that there is in fact no over-riding commitment to year-on-year stability. This is not because the Chinese do not care about 'standards': on the contrary. But it is worth emphasising, at this point, the multiple meanings attached to the word, and especially the difference between two of the dictionary definitions. One is that a standard is 'something set up and established by authority *as a rule for the measure of quantity, weight, extent, value or quality*'. However, another notes that standards may mean 'something established by authority, custom or general consent *as a model or example*' (Websters New Collegiate Dictionary: italics ours).

The Chinese commitment is less to the idea of standards as a measuring tool than to standards as an example and ideal. Their

purpose is not to create a benchmark and guarantee a lack of change but rather to encourage and provide incentives for continuous improvement. If more and more students reach the old pass mark for a key school or university, this is not seen as a problem, or a *prima facie* indication that the paper-setters have got something wrong: but rather as a response by the student population, reflecting hard work and achievement, which is to be welcomed and encouraged. The pass mark can be raised, and the exams will have achieved one of their purposes in providing a model and an example to strive after. The same principle justifies the whole key school approach, whereby favoured schools receive greater resources as well as the highest-achieving students. The authorities 'give them more and better resources in order to allow them to become 'models' for ordinary schools. Then the experience of key schools can be extended to improve ordinary schools' (Wang Gang 1995: 2).[2] Compared to this (generally accepted) objective, issues of equality of access are irrelevant.

United States

In some respects, China and the United States provide a dramatic contrast. The rhetoric of the one is as inclusive and egalitarian as the other—post-Cultural Revolution—is élitist. The American ideal is the comprehensive high school, not the selective key school; and political and legal action for over four decades has been devoted—albeit with very little success—to equalising the expenditures and quality of education available to all children in the public (state) schools. As a country whose per capita income is approximately fifteen times that of China, it can also afford a vast tertiary sector, in which anyone who wants to study can find a place, and where over a third of the working age population has gained some form of tertiary qualification. At the same time, this sector is similar to China's in one extremely important respect. It is highly diverse and highly hierarchical, in terms of entry requirements, and the perceived value of qualifications received. No-one pretends that

[2] Although the Chinese have participated in IEA studies, they have not provided anything that could be seen as a representative sample of students allowing for full comparisons with other countries. Most of the participating students have come from favoured East Coast schools, and American critics have seen this as a sign that the Chinese are trying to 'cheat' and get better results than they should. But from the Chinese point of view, the interesting question *is* whether their best students can 'beat' the rest of the world (to which these results will provide an answer).

'standards' are the same across the sector, or believes that it would be sensible, let alone practicable, for them to be so.

As most people know, the United States is also very distinctive in the nature of its educational assessment. The majority of tests used for high stakes purposes are developed and sold by independent companies, not by government-run or government-regulated agencies. Seven companies dominate testing, and in some sectors, just one or two companies do so—notably ETS and ACT at college level. (Fremer 1989) A large and growing number of commercially produced tests are used within elementary and secondary schools as well: for example, from 1970 to 1991 ETS revenues showed a compound annual growth rate of 11 per cent (Madaus and Raczek 1996). This growth partly reflects increasing use of tests by teachers for internal purposes (diagnosis, promotion decisions, information for parents) and partly increased demands by school district administrators and politicians for testing, in order to increase the 'accountability' of schools, and, supposedly, improve performance.

In spite of the explosion of test use, American students are only very rarely in a position where results on these tests have high stakes consequences. A very large part of the marking, grading, and certification that takes place in American education is completely separate from the testing industry's activities. High school diplomas are given essentially on the basis of grades awarded by class teachers, who *may* take some notice of test results (and often do not: Firestone 1998). Although many states are now introducing state-wide tests which must be passed in order to graduate, these are minimum-competency tests, and as such, low-stakes for most candidates. For the majority, their Grade Point Average is far more important than simple acquisition of a high school diploma; what it records is the average of entirely teacher-given grades. College degrees are gained, again, on the basis of teacher-awarded grades, obtained on a course by course (module by module) basis. In all these contexts, the English observer is struck by the teacher's total autonomy, and by the absence of moderation, quality checks, or even the most minimal form of double marking.

Not surprisingly, this creates problems of interpretation for those involved in large-scale selection activities. Selectors take it as given that standards (and so the meaning of a Grade Point Average) will vary from high school to high school, college to college. This may not matter if you are a small-town employer selecting from one or two schools' graduates: equally, at national level, most people will recognise the best and most prestigious universities' names and weight their ranking and

decisions accordingly. However, large-scale selection processes, such as characterise entry into undergraduate and graduate programmes in higher education, demand a simple, robust metric if they are to be practical and cost-effective. They also need to be defensible as 'objective'—especially in the United States, where legislative challenges play the regulatory role allocated elsewhere to national ministries and government agencies.

As a result, two tests—the SAT (Scholastic Aptitude Test) at college entry and the GRE (Graduate Record Examination) at graduate school entry—have become enormously important, high-stakes tests for young people seeking entry to higher education courses. Smaller, but equally high-stakes tests are important in specialised areas (e.g. entry to medical or law school); and there has been rapid growth in the subject-based Advanced Placement tests which give university credit (and also help win admission to the most selective schools, even where they will not give formal course credit against them). The SAT, GRE and related tests have been subject to consistent criticism on a number of counts, mostly from the professional assessment community, but also for their possible 'adverse impact' on particular groups of candidates. However, their importance has been increasing rather than decreasing; they have repeatedly met the key requirement, within American society, of satisfying the courts' criteria for an acceptable selection mechanism.

The development of American tests follows very well-established and enduring procedures. Although there has, in recent years, been a strong interest in breaking away from this, for example by developing 'authentic assessments', 'portfolios' and the like, these account for a tiny portion of testing and test development activity. (Koretz, Broadfoot and Wolf, 1998) The bulk remains firmly in the psychometric tradition, and produces tests which are commonly referred to as 'standardised' and/or 'objective'.

To describe a test as standardised is not really to do any more than indicate that scores have been transformed to fit a common metric—thus, the whole Japanese population is by now entirely used to expressing and discussing pupils' exam results in terms of 'Standard Deviation scores' (proportion of a SD above or below the mean). However, the term has also become associated, inside and outside the USA, with a particular approach and format: with the use of machine-markable multiple choice items, not tied to a particular school district or state syllabus, selected from a larger group of trialled questions on the basis of the way the various items 'behaved' during trialling. US tests are

generally normed on a large sample of the target population, meaning that results can be reported in terms of a national percentile rank or other norm-referenced scale, so that one knows how a given candidate scored relative to the national population. The SATs in particular, however, are also reported in the form of actual scores.

There is an enormous literature on the underlying assumptions or 'theories' of psychometric test construction (which also dominate the development of international surveys, including the IEA surveys and those of the OECD). I will make no attempt o summarise them here (though interested readers are referred to Wood 1991, Hambleton and Zaal 1991, Goldstein and Wood 1989, Goldstein 1996) However, the testing industry in its present form has two consequences which are highly relevant to the current topic. First, judgements about content and item format, and therefore about what a given level of success actually involves or means, are completely buried from sight. They take place at a quite early stage of a test's development, with far less scrutiny than is later given to the actual selection of test items, and with no public visibility or involvement at all. Secondly, partly because of the apparent statistical complexity and difficulty of test construction, there is a high level of public confidence in the objectivity of the tests, and in the presence of procedures which ensure comparability from one year and one test form to another.

The main preoccupation of the testing industry is not with ensuring that tests and items are equally difficult in some absolute sense, or sample exactly the same content or skills, but on whether alternative forms would produce the same *rank ordering* of individuals. If the main concern of the test industry was with some sort of absolute standard that students were supposed to achieve (as seems to be the case with the current governmental conception of National Curriculum levels); or with the effects of test items on the school curriculum, this might be a problem. But in the US context it is not. Over any short period of time, comparability and substantive standards are not actually very important, since the function of these high-stakes tests *is* so overwhelmingly one of ranking and selection.

It follows that, while 'standards' certainly are an issue in American politics, comparability of test standards, year on year, has not been. However, what the US shares with the UK is a culture of political suspicion of educators, a conviction that the country's education system is uniquely poor because teachers are no good, and a faith that political reform (including use of supposedly 'objective' tests for accountability

purposes) can help. This conviction has been fuelled by a number of specific events. One was the decline in SAT scores during the 1970s, which occurred fast enough to suggest a real decline in achievement (and fast enough not to be obscured by the regular re-norming of the test), and was never fully explained.[3] Another has been the recent performance of American children on international mathematics surveys (including the IEA's maths and science surveys, SIMSS and TIMSS).

In response to the TIMSS results, the US federal government is encouraging states and districts to use the TIMSS tests with their eighth grade students as part of a 'Benchmarking' exercise. The idea is to use the tests as a way of developing fixed benchmarks or 'performance standards': i.e. standards in the sense of a 'rule for the measure of quantity . . .' Such a notion has not been central to the functioning of American education in the past, and, unless there are enormous changes in selection procedures for higher education and the professions, it seems unlikely ever to have the force the Feds desire. A commitment to local and state control of education predisposes politicians and educators alike to reject any central yardstick. The size and the complexity of the sorting process between high school and college, and between undergraduate and graduate programmes are also crucial. They militate against any uniform and simple set of 'standards' becoming a basis for certification or selection.

Sweden

To move from the US to Sweden is to turn from a country whose education appears test-dominated to one where testing seems invisible. Sweden does not fit either of the dominant models outlined above: it has no examination boards or public examinations, and no widely-used standardised tests. It does, however, participate in the international tests—IEA and IALS for example—and takes them very seriously.

For a small country, Sweden has also received an unusual amount of attention over the last 50 years, especially from economists and political scientists. Its combination of extremely wide-ranging welfare state provision and high tax rates with a highly successful economy, encompassing a number of world-class companies (Ericsson, ABB, Volvo etc),

[3] In spite of extensive and impassioned debate, and copious research, it remains unclear whether there was any major downturn in achievement, or why. See e.g. Stedman (1998).

has led many commentators to seek both insights and recommendations for their home countries. In recent years the 'Swedish model' has been less generally successful, and subject to criticism and modification at home: but the country's social policies and ethos remain highly distinctive.

Many educationalists, if asked what they knew about Swedish education, would reply that the country has no external public examinations. This is, in fact, only half true. What is true is that, like the United States, Sweden has no formal final examinations on which leaving certificates or university entrance are based. However, there *are* national tests, used at key points during secondary education; and their nature, purpose and history indicate a great deal about the Swedish approach to 'standards'.

As in many other countries, Swedish education and assessment are currently in the process of reform. The changes will affect the nature of the national tests, but at the moment they are still at a trial stage (and encountering serious implementation problems). The final shape of the reforms is not yet clear, and the discussion here largely describes the 'old' system; but the basic purpose of the tests, and the role of teacher assessment will in any case remain unchanged.[4]

There are two major points of assessment which affect a Swedish pupil's future path. The first is at the end of ninth grade, when all subjects that the pupil is studying are assessed by the relevant subject teacher, and given a mark: historically from 1 (lowest) to 5 (highest) although under current reforms this is being changed to a four-level scale. The number of subjects could total as many as 15 or 16 subjects, and all subjects count towards a student's final assessment and set of marks, which is simply an unweighted average. The second major assessment is at the end of twelfth grade when the same procedures apply.

The assessment at the end of compulsory school (9th grade) determines admission to different courses within the upper secondary school, or (for a decreasing number) affects labour market entry; and the assessment at the end of 12th grade determines admission to university, or, again, is important in making job applications. In 12th as in 9th grade, up to 15 subjects are taken and all count equally. (Thus the mark for a one semester course in child care could in theory count

[4] A major purpose of the reforms is to make tests and teacher assessments alike more criterion-based and less norm-referenced.

for as much as the maths mark or Swedish grade. In practice, everyone gets a more or less equivalent mark for this course.) Almost all Swedish students now proceed directly to upper secondary school after 9th grade. Here, recent reforms have made the vocational and academic options in upper secondary school the same length (three years) and much more similar than in the past. However there is still a clear distinction between them, related (among other things) both to the school marks required for admission, and to the university programmes whose prerequisites they meet.

Admission to both upper secondary programmes and university is made on the basis of the marks awarded by teachers (using a 1 to 5 scale) at the end of the 9th and 12th grades respectively. However, a number of national tests exist which are designed to ensure comparability of standards across the country. Under current arrangements the first national test of students take place in the first semester of grade 8. This covers English, and is followed by tests of Maths and Swedish in grade 9 (the end of compulsory school).

The tests are written by university-based groups in education faculties, who draw on the national curriculum and their own experience to prepare the items. They are taken by all students, and are used by teachers to standardise their own marks. However they do *not* override them. They work as follows.

All students take the tests, and a large sample of scripts is marked and analysed centrally, providing a mark distribution which can be divided up into 5 bands equivalent to the grading 1 through 5. This corresponds to the 5 level scale used for the official assessment by the teachers. (The use of a five-level scale dates back to 1962. Before that, 7 levels were used.)

The bands or scale followed the normal distribution quite strictly in the early days, so that the percentage of pupils falling into each band nationally was as follows (Kilpatrick and Johansson 1994):

Band	1	2	3	4	5
Percentage of Pupils	7	24	38	24	7

More recently the norming system for grade 9 was changed. Guidance states that the mean should be 3 for the country as a whole, and that there should be more 4s and 2s in a *class* (sic) than 5s and 1s respectively (Skolöverstyrelsen 1980: quoted in Kilpatrick and Johansson. See also Wolf and Steedman 1998) Nationally, about 40 per cent do indeed get a 3; 30 per cent get a 1 or a 2, and 30 per cent a 4 or a 5. No

information is compiled on the breakdown between 1s and 2s and 4s and 5s. However, given that there was a lot of unhappiness about giving low grades—particularly 1s—and that this affected the move away from the standard grade distributions of the 1960s, it seems likely that there are fewer 1s than a normal distribution would imply.

The purpose of the national tests is, as noted above, to help the teachers standardise their own marking. The teachers mark the tests themselves, and so know exactly what raw scores each child has obtained; but *there is no direct link between an individual child's test and end-of-year score.* Instead, the information from the national analysis tells a school what proportion of its students should, at the end of the day, fall into a given category. So if, for example, the marks which students get on the test indicate that 40 per cent fall in band 3, then the teachers and school must stick to that in their final grades. Their distribution of grades for, say, Maths or Swedish—the standardised subjects—should provide for 40 per cent receiving grade 3s, even though the individual students getting a 3 at year end may not be the same as the ones who got marks in the grade 3 band on the test itself.

In the past, teachers were told which raw test scores corresponded to which grades for all 5 grades, and so could align their mark distribution exactly with that suggested by the national test results. More recently, in line with the more permissive (some would say ambiguous) marking guidance, they have been told only which set of scores covered the middle grade (grade 3). No guidance is given on where the mark cut-off between grade 2 and grade 1 lies, or the one between 4 and 5. If 30 per cent of the school's cohort scored higher than the grade 3 band of marks, it is up to the school where it puts the higher boundary, and how it allocates that 30 per cent to the two grades.[5]

This process is quite straightforward for teachers in subjects with a national test. In others they have to use other methods—e.g. comparing science results with those in maths, on the assumption there should be some relationship. There is group discussion of the grades in many schools; and the head, who has ultimate responsibility, will ask for grades to be justified.

The remarkable aspect of the Swedish approach is the way it lodges final judgements with teachers, as those best placed to know a pupil's performance; but uses testing not to replace but to improve

[5] In fact in Maths the way results are reported makes it possible for teachers to calculate the marks for each band quite easily, should they wish.

that judgement. For the British outsider, however, it is natural to ask how this system can actually deliver genuinely consistent judgements, especially in the subjects not covered by the tests; and how it can survive given the extremely high stakes nature of some of the judgements made. At entry to upper secondary school, where 9th grade marks determine which upper secondary programme a student can enter, there is considerable room for flexibility, discussion with the students and parents, and, indeed, school autonomy in deciding whether or not to admit a student to the desired course. At university level, however, no such flexibility is found. Entry to Swedish university courses is strictly on the basis of marks: candidates are ranked, and places are allocated in strict descending order. If one place is left, and there is a difference of 0.1 in the average marks of the two remaining applicants, the place goes automatically to the higher scoring candidate.[6] To make allowances for the fact that a student comes from a poor background, or a small rural school, or has faced family problems in the preceding year—as Oxford or Cambridge selectors do routinely—is completely outside the powers of a comparably important faculty at the Universities of Stockholm, Uppsala or Gothenburg.

In the Chinese case, we have seen how the main objective of the university entrance procedure is to secure and improve the level achieved by the highest achievers, and to identify them in a 'fair' way, rather than to attempt to identify underlying aptitudes, let alone use the selection system to redress (even partially) underlying inequalities of opportunity. In the Swedish case, however, this interpretation seems implausible, and is, indeed, completely at odds with the egalitarian and inclusive philosophy which, still, marks out Swedish culture and politics. How is it then, that the education system and society at large should evince such a stable and long-standing lack of concern over whether standards are equal between schools, or subjects, or years?

The answer is, I would suggest, to be found in a different aspect of that same culture, which, in this case, takes precedence: a belief in consensus and a corresponding tendency to respect and indeed build up the power of professionals and professional groups, rather than seeing them as essentially a conspiracy against the public interest. This is a theme to which I shall be returning in some detail at the

[6] There is, in fact, since 1991 a fail-safe mechanism in the form of an aptitude test which students can take; and which may be used partially to substitute for school marks.

end of this paper; but in the Swedish context, it means that education policy is conducted with rather than against the teachers, who value and defend their traditional role as student assessors, and who are seen as the best-placed to deliver those judgements. None of the Swedish education reforms of the last fifteen years, which include efforts by the conservative government of the early 1990s to encourage independent schools, competition between universities, and closer links to industry, have involved any significant questioning of the teachers' position as assessors and so promulgators of 'standards'. Standards are not an issue because the competence of their guardians is not an issue either; and because standards are also seen as a dimension of people's expertise, not as embodied in external mechanisms and instruments.

France and Germany

The final ports of call on this rather odd Cooks tour are our two largest European neighbours, France and Germany. Very different from each other in all but two crucial respects, their examination systems bear a close resemblance to, respectively, those of China and Sweden; and will therefore be described relatively briefly.

France possesses a large number of recognised national diplomas awarded outside the universities, all of them assessed entirely or in large part through public examination: of which the baccalauréat is the largest and most important. The papers for these examinations are set centrally, by the Ministry of Education, using questions selected from among large numbers contributed by the regional offices—the *Rectorats,* one per *Académie*—which have, among their duties, the actual conduct of the examinations. This process is taken extremely seriously, in the sense that the papers—taken at exactly the same time, under classic 'exam conditions', by all candidates—are published and distributed under conditions of extreme confidentiality; marked under the supervision of the *Rectorat* officials; and the results ratified by an officially constituted *jury.*

This process is regarded as central to establishing and maintaining an objective and fair system, in which candidates succeed on their merits. While French commentators, most notably Pierre Bourdieu (1989), have subjected this belief in the system's fairness to sustained criticism, it remains strong. The continuing commitment of the French élite to centralised control over education is itself linked to a conviction that only in this way can one sustain a system of national diplomas: that

is, diplomas which are recognised as objective, consistent, and equally valuable whoever the holder, and wherever and whenever they were obtained. However, by English or American standards, the actual process of examination setting and marking is remarkable for its almost total neglect of 'technical' procedures such as item scrutiny, comparability exercises to check inter-*Académie* marking consistency, archiving of scripts etc. There is fairly little checking of teacher-markers' scoring, and no national appeal procedure. Instead, it is assumed that, on substantive issues relating to standards, the professional expertise of those setting the questions will ensure acceptable levels of consistency and curriculum fidelity.

The German system is even more free of technical procedures but in other ways a total contrast to the French. The German secondary school system is (broadly) tripartite, with vocational, technical and academic schools (*Hauptschulen, Realschulen, Gymnasien*), each of which leads to a school leaving certificate which is gained on the basis of examinations. Only the certificate from a *Gymnasium*—the *Abitur*—allows entry to university, although the *Realschulen* certificates allow holders to enter polytechnic (*Fachhochschulen*) courses.

In every case, the school leaving certificate is awarded on the basis of internal examinations and marks, awarded by students' own teachers. The expected standards are defined in written documents produced by the standing conference of state ministers (since education, in Germany, is a state not a federal responsibility.) In some states (*Länder*) the topics for the written examinations are set at state level; elsewhere, schools submit topics which must be approved (to establish common expectations across schools) but are their own responsibility. Marking is entirely internal, with the school principal having the responsibility to maintain and monitor standards.

I noted above that these apparently quite disparate systems of France and Germany do have two important characteristics in common. The first of these is that, unlike England, both have experienced an enormous increase in numbers of candidates and diploma-holders without also experiencing serious loss of confidence in the system. In France, a government commitment to have 80 per cent of the age cohort completing baccalauréat programmes (though not necessarily passing the exams) has almost been realised, involving more than a doubling of participation rates since 1985. In Germany, the percentage obtaining the *Abitur* has risen from 5 per cent in 1960 to 25 per cent in the early 80s to over a third today.

There are, of course, some concerns about resulting achievement levels. In France, a good part of the increase in baccalauréat student numbers is found in programmes leading to the vocational bac (*baccal auréat professionel*) rather than the traditional academic courses. These vocational baccalauréats are seen as definitely of lower status, and few holders go on to university. But a very large part of the increase has been in 'general' bac. numbers, and here, especially when compared to English worries over A level standards, confidence has by and large been maintained. A similar situation obtains in Germany. There are worries about comparability of standards, but these centre around the conviction of the Bavarians that all the other *Länder* are operating with lower standards than they are. Moreover, this is not a serious issue; threats by the Bavarians to refuse to allow *Abitur*-holders from other states automatic entrance to Bavarian universities have come to nothing, and the German population generally shows a continuing confidence in both the quality of its schools and the reliability and worth of their certificates.

The other major similarity between the two countries is their reliance on professional judgement and expertise, and corresponding absence of any widespread culture of assessment expertise, concerns with statistically based quality control, and the like. I will not speculate on how far this absence is a cause rather than a result of politicians' willingness to accept professional control, and to leave the general conduct of educational policy—whether reformist as in France, or highly stable as in Germany—to their officials and teachers. However, it is worth noting how differently educational interest and identity group boundaries run, as compared to England (but perhaps not Scotland.)

Far from seeing the *Académies* as rivals for power, as English ministries and quangos see the local authorities and exam boards, the French operate a centralised system in which the senior officials at the Ministry of Education form part of the same group as those who run the regions and the examinations. In Germany, teachers are civil servants, and enjoy both the respect and the protection this implies. Challenging, or allowing challenges to, their assessment authority would call into question the position of public servants throughout the educational (and other) spheres. In the absence of any general concern over standards, it is thus not surprising that neither country's politicians has shown the slightest inclination to open up the assessment process to public scrutiny, by, for example, instituting rights of appeal,

oversight committees or sponsoring large-scale research studies on comparability-related issues.[7]

However, there is also a link between levels of public concern over standards and the role those standards play in determining students' lives. While both France and Germany use school-leaving examinations to determine university entrance, their higher education systems are very different from the highly hierarchical and internally differentiated Anglo-Saxon (or Chinese) ones. Leaving aside the French *Grandes Écoles*, which operate their own formal entrance examinations, both countries essentially allow university entrance to all holders of the relevant diploma. Thus, in Germany, an *Abitur* holder can study any degree in any university in the country. (The only major exception is medicine, where the *numerus clausus* system holds, and a tiny fraction of a mark can make the difference between success and failure to win a place. Significantly, here there are alternative routes and fail-safe procedures which allocate about 20 per cent of the places differently.) In France, there are tighter requirements for prerequisite subjects: but again, anyone with a 'general' (academic) bac. can be sure of a place. Most students study locally: and the important thing is that one passes one's bac, rather than how well one passes. In consequence, while the certificates are high stakes, it is only one boundary (pass/fail) that really matters. It is not that important to most students' chances whether they got precisely the right mark. In other words, compared to England, *only a small proportion of students have a strong, personal interest in the precision of the marking*, and they are not the most successful or visible ones. This, too, tends to reduce the pressure on the system to deliver apparent comparability within and across years; and also reduces the extent to which standards become a public and political issue.

Conclusions

The limited number of ways in which countries organise high-stakes examinations and diplomas suggests an equally common approach to maintaining 'standards'. The previous discussion will hopefully have indicated that this is not the case. In fact, the way countries operate,

[7] The recent refusal of the French to allow publication of the French results from the International Adult Literacy Survey may reflect not merely a refusal to countenance, let alone admit, that French adults could be performing poorly, but also concerns over the public impact of any such findings.

and the degree to which aspects of 'standards' become contentious, can only really be understood in terms of far wider aspects of political and organisational cultures. It is not because they operate with such different assessment systems that the United States and Germany are hugely different in the extent of political concern over educational standards. Rather, in each case, the assessment and certification systems must be understood in terms of higher education structures, the role of legislation and litigation, and deeply embedded attitudes towards the position of public servants (including teachers), and the efficacy and respect due to regulations.

A number of other points suggest themselves. The first is that *England—and indeed the UK—is extremely unusual in its overt claim that 'standards' are being maintained from year to year in some absolute sense*, and in the primacy of criterion-related concerns over norm-referencing practices. Other countries may make the implicit assumption that 'standards' are being held constant, in the sense of some measure which yields the same quantity year or year; but they certainly do not make models of performance or notions of benchmarking the centrepiece of their item-writing, examining and moderating procedures in the way the UK examination boards have done for many decades. Conversely, other countries tend to be much more overtly focused on the process of differentiation and selection, through a more or less explicit norm-referencing approach. It is commonplace to argue that certificates such as A levels have the dual function of certification and selection, and that the latter function is dominant. However, compared to most countries, what is striking about the UK is that we tend to spend most of our effort on procedures and concerns which are more relevant to certification than to selection pure and simple.

The second point which strikes one is that transparency and public confidence are by no means positively correlated—perhaps the opposite. In countries where the assessment process is left very much to the teachers and educational professionals, there appears to be less, not more anxiety over standards. It is always difficult to know which way lines of causality run, and the increasing extent to which the examination process in the UK is overseen, dissected, and opened up to appeal is certainly in part a response to, rather than a cause of, public unease (Wolf 1995). But given the inherently imperfect and non-mechanical nature of assessment judgements, and the UK's unusually ambitious claims regarding standards over time, one may doubt that further

auditing and oversight will increase confidence, and predict that it is more likely to decrease it further.

The third and final point follows on from this. Whichever country one looks at, it is clear that, ultimately, 'standards' do depend on professional judgement. From an English standpoint, we may be somewhat unimpressed by French or German neglect of comparability studies and agreement trials, or by their level of attention to checking marks and markers; or see much of the underlying 'theory' in US psychometrics as a combination of the doubtful and the true-by-definition. Nonetheless, at root, all of these countries, like our own, build their systems on the foundation of professional knowledge and professional judgement about what to ask, and how to assess the response.

This irreducible fact runs counter to a set of beliefs which are especially powerful at present in Anglo-Saxon societies but increasingly influential elsewhere as well, and which demand 'accountability', 'openness' and demonstrated 'value for money'. In a recent book called *The Audit Society: Rituals of Verification*, the LSE's Michael Power has described in detail how pervasive these ideas have become, not just in education but in the whole area of regulation within organisations as well as across sectors. Unfortunately, in almost every case, whether one is talking about the auditing of a company; new management techniques for the Health Service; the quality management approaches being introduced into higher education across Europe; or the growing use of tests by British and American politicians, it is extremely unclear just what is being measured, or what can be done with the information. Things get counted because things have been found or created which are countable: as Power puts it, 'audits work because organisations have literally been made auditable; audit demands the environment, in the form of systems, and performance measures, which make a certain type of verification possible' (1997: 91).

This movement is in part a result of a messianic belief in the power of these approaches to make things 'better': more efficient, fairer, more productive. However, they are also a result of a breakdown in trust, and especially in trust in professional groups. It is not clear to me how far this is a cumulative and more or less inevitable component of all modern societies, as I suppose some sociologists such as Habermas would argue; and how far it is associated with more specific movements and analyses, such as the distrust of all organised groups, as conspiracies against the public, which marked the Thatcher years.

Either way, an absence of trust is not really sustainable, because,

while trust is always fragile, it is also completely necessary to the functioning of social life, and certainly to the operation of any system of educational certification and selection. If we cannot restore and improve public and political trust in our current 'standards' there is a risk that we will follow the American example and be forced into restoring confidence by establishing new and totally *non*-transparent 'expert systems', dominated by issues of reliability. We would, in the process, end up with an increasingly constrained form of public examination whose apparently 'objective' nature merely hides the value judgements and decisions inherent in any assessment. It would be far more desirable, in my view, to learn from other countries that effective learning and high standards do not require a supposedly 'ever-fixèd mark'; something that in education (if not in love) is in any case quite unattainable.

Discussion

Julia Whitburn

Professor Wolf has provided us with an admirably clear exposition to support her thesis that there is little international consensus regarding methods of assessment. She has also argued that the lack of consensus reflects differences in basic values, beliefs and educational objectives of the countries concerned. I do not wish to dispute her main argument but rather to offer a few comments on the value of a comparative perspective with regard to the issue of educational standards.

The significance of a comparative perspective is perhaps most immediately apparent in the context of the large international studies of achievement, notably those by the International Association for the evaluation of Educational Achievement (IEA) and the most recent Third International Mathematics and Science Study (TIMSS). Despite the well-documented difficulties of making valid trans-national comparisons, such studies are here to stay, and it is only fair to say that many of the methodological deficiencies of the First International Mathematics Study (FIMS) were addressed by the time of the third study. Lessons to be learned from these studies fall, I would suggest, into three categories. First, we see, albeit crudely, how we stand in the international league table of results. This gives our national pride, which

thrives on the spirit of competition and knowing how our achievements stand in relation to those of others, a blow when we look at mathematics attainments but a boost when we quickly turn to the science results.

The second lesson, however, is perhaps of greater importance: we learn that simplistic comparisons of educational systems and results are both misleading and dangerous, and that the multi-factorial and complex nature of the educational arena is aggravated by the cultural contextualisation. In spite of this, however, as Professor Wolf has indicated, countries do 'sign up' to the proposition that there is an 'educational construct' that has a commonality of meaning and relevance for all countries involved and that this can be quantified in an adequately valid way.

The third lesson to be learned from international studies comes in our attempt to turn from looking outwards to the achievements of other countries to looking inwards to our own achievements. International studies help to ensure that the debate over performance standards is informed by our understanding of the details of educational systems and circumstances prevailing in other countries, and lead us to ask ourselves questions about the *structure* of our own educational system, which we might otherwise not have asked. This third lesson, I believe, is the one that, in any educational context, represents the real value of a comparative perspective.

Professor Wolf has, in her paper, interpreted educational standards as the measuring of achievement by pupils within a single country and she has drawn to our attention the different and contrasting approaches to assessment or output measures. The real question is 'What does this comparative dimension add to our understanding of, or debate on, measuring achievement?'

First, we become aware that if we are to judge by the amount of testing, in England there is a great concern with standards. English pupils are tested on a nation-wide basis more frequently than pupils in many other countries. Some countries have no system of national testing of all pupils at any stage, for example, in Japan, which is one of the highest achieving countries in terms of average mathematical attainment. In Switzerland another high-attaining country mathematically there is no point during the period of compulsory formal schooling at which all pupils are routinely assessed using a standardised test, although the highest attaining pupils attending *Gymnasien* may obtain the *Maturität* or school-leaving examination equivalent perhaps to that of the *Abitur* (of the better schools) in Germany. In contrast, *all* pupils

in England are tested by nationally standard tests at ages 7, 11, 14 and 16, at the end of each Key Stage during compulsory schooling. In addition, a national system of Baseline Assessment of all children on entry to schooling has just been introduced. Increasingly, schools choose to administer additional School Assessment Tests (SATs) tests between KS1 and KS2, and, on leaving school, a high proportion of pupils also take 'A' level examinations or other forms of assessment at 18.

A comparative perspective makes us consider more carefully what the *purpose* of this testing is and to ask whether or not it could be achieved better in other ways. At 16 and at 18, testing may be justified on the basis of certification of individual pupils and/or selection for employment or the next stage of education (the latter reason is probably more relevant to 'A' level assessment). Testing at Key Stages 1, 2 and 3, however, is unrelated to the certification or selection of individual pupils. Indeed, results relating to the performance of individual pupils are available to parents only in the most general form in terms of *level* gained, defined very broadly.

While it must be a source of reassurance for parents to know that their child has achieved at least the so-called 'expected' level 2 at age 7, this is scarcely a reason for congratulation. In 1997, for example, the level was achieved by 84% of children. (Schools may inform parents as to whether their 7-year-old child got a Level 2c, 2b or 2a, but the amount of explanation provided varies from school to school.) In fact, analysis of SATs results tends to focus more on *results at school levels* and for *schools* to be congratulated on the improvement to their achievements. (This is also now true of GCSE and 'A' Level results, where league tables of *school* results highlight the best and worst.) We are all aware that making simplistic comparisons of school achievements can be misleading.

Indeed, these are arguably more dangerous than simplistic *transnational* comparisons since we are not able to perform the multifactorial analysis of school situations that we recognise as necessary for valid international comparisons. But why is it that in England our obsession is with comparing *school* performance? Does it reflect our mistrust of our schools and our teachers which has been fuelled by those in influential positions? Or is it more a reflection of our unwillingness to attribute individual responsibility for achievement (or failure)? In Japan, there is a widespread belief in the importance of effort rather than innate ability and *pupils* are encouraged that 'If you work

hard enough and persevere, you can succeed.' In England, I would suggest, the message is that *teachers* need to work harder and persevere in order for their pupils to succeed and where pupils do not achieve well, it is poor teaching that is held to be responsible. From the view that teachers and schools are to be blamed for their pupils' poor achievements it is only a small step to the system of payment by results which operated in schools over 100 years ago, as described by Professor Aldrich. Indeed, there are already financial consequences for low-achieving schools since lower enrolment of pupils which can follow publication of their poor league table rankings then adversely affects their subsequent funding levels. Incidentally, if the purpose of national tests is to monitor school rather than pupil performance, it is possible that this could be achieved effectively and at a lower cost by using sampling rather as the APU did at an earlier time.

But that is a digression. If our tests are to be about school rather than individual achievement, this suggests that they may be moving towards a concept of 'expected standards' as outlined by Professor Aldrich. We need to be very careful, however as Sig Prais has already explained in our use of the term 'expected standards'. The statement that by 2002 '75% of 11-year-olds will be reaching the standards expected for their age in maths' implies a curious distribution of mathematics scores. What, perhaps, the government hopes to achieve by 2002 is for 75% of 11-year-olds to be achieving a specified *minimum* standard.

This brings me to the third question that benefits from a comparative perspective, namely, what can we learn in relation to the concept of minimum standards by considering other educational systems? The concept of a minimum standard for a particular grade level is not uncommon and may be found in other countries such as France, Switzerland and Germany. In each of these countries, a pupil's progress to a subsequent grade is contingent on his/her satisfactorily achieving the minimum standard. What we find, however, is that to enable all pupils to achieve the minimum standards, certain modifications or additions to educational systems develop. For example, we have the practice of '*redoublement*' in France, known as '*sitzen bleiben*' in Germany, and '*repetieren*' in Switzerland in conjunction there with the more important element of age-flexibility on entry to schooling. In Japan, the private tutorial system of '*juku*' is an essential underpinning for the public education system in which minimum standards are expected to be attained by all pupils. Without these practices, it is

difficult to imagine that the expectation of minimum standards for all pupils could continue.

If the government is indeed proposing that 75% of all 11-year-olds should achieve a *minimum* standard, we need to ask what this standard will be, how it will be decided and what the implications are for the curriculum and for the organisation of classes and of teaching. We may wish to consider how our concept of a minimum standard compares with those of other countries. For example, if we look at the content of the Maths curriculum for elementary school children in Japan, we might conclude that the content and level of the curriculum for children aged 7 are not dissimilar to those of English pupils. If 84% of pupils achieve a Level 2, are they then not performing at a comparable level to 7-year-old pupils in Japan? All research evidence points to the fact that English pupils do *not* perform at a level comparable to Japanese pupils at this age. One reason for this apparent paradox may lie with the way in which the 'standard' required to achieve a Level 2 is set. Many of us may be familiar with the content of the tests; some may *not* dispute the *level* of *difficulty* of the questions. But are we all aware, and are we content that the *level* of the expected standard (or minimum standard) is relatively low; for example, this year for children to be awarded the 'expected' Level 2, the KS1 Mathematics test required a score of only 10 out of 36 and a similar percentage score for science?

I have mentioned only a few of the questions that we might ask in relation to educational standards: a comparative perspective helps us to focus our minds on these questions that we need to be asking in order to ensure that what is being promulgated as an 'educational standard' is of value. If we are going down on the route of minimum standards then we need to be sure we understand and agree with what that is. Pupils, parents and the country deserve this.

David Reynolds

Alison Wolf's paper makes some very important points. She notes that the apparent universality of concern about standards and the increasing international participation in surveys of achievement, and indeed the whole discourse about the results of these surveys, hide large differences between countries in how 'standards' are maintained, understood and operated.

I would venture three comments. First, I think that the paper under-

estimates the extent to which there is not just considerable variation between societies in their systems for ensuring standards but also in their very valuation of which 'standards' are important. It is a much neglected feature of the 'tyranny of the international horse race' that the countries who have historically done well on the various academic outcomes, measured conventionally on tests of basic skills, are now most doubtful of their value. In anglophone societies, such as the United States in the case of its maths performance at 16, and England and Wales in the case of its maths performance at 9, there is considerable attention given to the surveys and a correspondingly increasingly narrow view that academic outcomes are important. Countries doing well, such as those of the Pacific Rim, are concerned with new definitions of what 'standards' matter. Instead of any national concern about academic standards, there is a concern:

- to broaden the range of children's capacities to include more higher order skills in addition to the basic skills that the system apparently transmits so effectively;
- to focus upon social outcomes that are important for the world of work, such as children's capacity to work together collaboratively in groups;
- to encourage children to be creative and produce 'new' knowledge, and to think laterally to generate the interactions between bodies of knowledge.

These concerns are now widespread across the Pacific Rim. I would expect this cross-national variation in what constitutes the 'important standards' to increase, notwithstanding the continued popularity of academically based international studies. I would expect consequently much change in assessment systems as new outcomes become standards.

Secondly, Professor Wolf's paper misses one important international characteristic of the standards debate, which is how 'standards' are increasingly being linked to specific 'policies' and not merely seen in terms of performance on outcome measures. It is one thing to propose, as some of us have done, that we should investigate the extent to which other societies have classroom characteristics which may be useful and effective if transplanted to our own society. An example of this is the interest in Pacific Rim whole class interactive teaching. However, there appears to be an international movement to look at and potentially to adopt the national *policies*.

All this may be very damaging, however, since it may be that *different* policies may be required in *different* national contexts to generate the effective classroom teachers that may be quite *similar* in different nations. In our International School Effectiveness Research Project (ISERP) study (Reynolds et al., 1998), we have found that certain teacher instructional features are effective across nine societies (as varied as Taiwan, the United Kingdom, the United States and Australia) in discriminating ineffective schools from typical schools and from effective schools. The clear demonstration of high expectations, frequent questioning, lesson structure, use of review, pupils' time on task and teachers' capacity to manage their classrooms are factors that all exist in the effective schools of different countries. However, the school level variables and the national policies that are necessary to generate these *same* characteristics are very *different*. In a country such as Taiwan, where there is a cultural acceptance of directive political leadership, a Principal may need to be directive with teachers to get these effective classroom factors in place, because they would not discover them for themselves. In a culture such as Great Britain, such direction may not be functional. At national policy level, to speculate with another example, the use of the 'meso' level of the District or the LEA to lever up standards of teaching may be more effective in cultures where there are traditions of decentralisation to federal governments than in the United Kingdom, with its strong central government.

As 'standards' become linked to 'policies' for the obvious political reason that policies are easier to explain than processes and because one can get changes in policies quicker than in outcomes, they are tending to become more similar. I suspect by contrast that a more fruitful perspective may be the maximisation of policy variation as the means to ensure the universalisation of the same effective teaching characteristics. In any event, the 'for standards read policies' movement has considerable implications.

Thirdly, I found the comments about our British near obsession with the accountability of the educational system as resulting from the breakdown of 'trust' between educational 'producers' and the wider society, most interesting. I suspect that this breakdown reflected public perception of the wide variation in school quality that exists in the United Kingdom, a variation which school effectiveness research revealed and which politicians then exploited for political gain.

I also suspect that British teachers themselves have made the breakdown of trust more serious by their own inability to trust each other in

close professional relationships. Perhaps reflecting high needs for personal autonomy, British teachers have been reluctant to permit others to see them teach, a reluctance magnified in its effects by the lack of time in which such observations can take place and the lack of observation systems of proven reliability, validity and practical applicability. More recently, the arrival of various forms of market pressures involving competition between schools and associated systems of parental choice has probably made any transfer of 'good practice' between schools as rare as such transfer between individuals.

Building on Professor Wolf's important insight, I would speculate that the way to rebuild trust by the public in education professionals is to deal with the levels of between school and within school variability, which themselves may be a direct result of the lack of trust amongst professionals in education.

Bibliography

Adams, J. (1912). *The Evolution of Educational Theory* (London, Macmillan).

AIE (1996). *Assessment in Education*, **3**(2).

Aldrich, R. (1995). *School and Society in Victorian Britain: Joseph Payne and the new world of education* (New York, Garland).

Aldrich, R. (1996). *Education for the Nation* (London, Cassell).

Aldrich, R. (1997). *The End of History and the Beginning of Education* (London, Institute of Education).

Aldrich, V. C. (1963). *Philosophy of Art* (Englewood Cliffs, Prentice-Hall).

Anderson, R. D. (1995). *Education and the Scottish People* (Oxford, Oxford University Press).

Arnold, Matthew (1863). *A French Eton*, reprinted in *The Complete Prose Works of Matthew Arnold vol. ii, Democratic Education* ed. R. H. Super (Ann Arbor, University of Michigan Press, 1962), pp 262–325.

Arnott, M. (1993). Thatcherism in Scotland: an Exploration of Educational Policy in the Secondary Sector (PhD Thesis, Strathclyde University).

Ayer, A. J. (1946). *Language Truth and Logic* Second edition (London; Penguin).

Baird, J. (1998). What's in a Name? Experiments with blind marking in A-level Examinations. *Educational Research*, **40**(2), 191–202.

Baird, J. and Jones, B. (1998). Statistical analyses of examination standards: better measures of the unquantifiable? (Associated Examining Board Research Report—RAC/780).

Bardell, G.; Fearnley, A. and Fowles, D. (1984). *The contribution of graded objectives schemes in Mathematics and French* (Manchester, Joint Matriculation Board).

Barnes, B. (1974). *Scientific Knowledge and Sociological Theory* (London, Routledge and Kegan Paul).

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd edition) (London, Arnold).

Bartholomew, D. J. (1996). *The Statistical Approach to Social Measurement* (San Diego, Academic Press).

Beardsley, M. C. (1981). *Aesthetics: Problems in the Philosophy of Criticism* (Indianapolis, Hackett).

Beaton, A. E. and Zwik, R. (1990). *Disentangling the NAEP 1985–86 reading anomaly.* (Princeton, Educational Testing Service).

Benn, C. and Chitty, C. (1996). *Thirty Years On* (London, David Fulton).

Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality* (London, Penguin).

Berry, C. (1997). *Social Theory of the Scottish Enlightenment* (Edinburgh, Edinburgh University Press).

Best, D. (1985). *Feeling and Reason in the Arts* (London, Allen & Unwin).

Bierhoff, H. (1996). Laying the foundation of numeracy: a comparison of primary

school textbooks in Britain, Germany and Switzerland. *Teaching Mathematics and Its Applications*, **15**, 141–60.

Billington, R. (1988). *Living Philosophy: An Introduction to Moral Thought* (London, Routledge).

Bourdieu, P. (1989). *La Noblesse d'État: Grandes Écoles et Esprit de Corps* (Paris, Les Editions de Minuit).

Brock, M.G. and Curthoys, M.C. (1998). (eds.). *The History of the University of Oxford vol. vi, Nineteenth-Century Oxford, Part 1* (Oxford, Clarendon Press).

Brooks, G. (1997). Trends in standards of literacy in the United Kingdom, 1948–1996 (paper presented at the UK Reading Association conference, University of Manchester, July 1997, and at the British Educational Research Association conference, University of York, September 1997).

Brown, A., McCrone, D., Paterson, L. and Surridge, P. (1998). *The Scottish Electorate* (London, Macmillan).

Burnhill, P., Garner, C. and McPherson, A. (1990). Parental education, social class and entry to higher education, 1976–1986. *Journal of the Royal Statistical Society*, series A, **153**, 233–248.

Burstein, J., Kaplan, R., Wolff, S., and Chi, L. (1997). Using Lexical Semantic Techniques to Classify Free-Responses (Princeton N.J. Educational Testing Service Research Report available on ETSnet at http: //www.ets.org/research/siglex.html).

Christie, T. and Forrest, G. M. (1981). *Defining Public Examination Standards* (London, Schools Council/Macmillan).

Cipolla, C. M. (1969). *Literacy and Development in the West* (London, Penguin).

Clanchy, M. (1979). *From Memory to Written Record: England 1066–1307* (London, Edward Arnold).

Collins, R. (1979). *The Credential Society* (New York, Academic Press).

Committee of Council on Education (1863). *Report of the Committee of Council on Education 1862–63* (London).

Committee of Council on Education (1872). *Report of the Committee of Council on Education 1871–72* (London).

Committee of Council on Education (1873). *Report of the Committee of Council on Education 1872–73* (London).

Committee of Council on Education (1883). *Report of the Committee of Council on Education 1882–83* (London).

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction* (Oxford, Blackwell).

Cox, C. B. and Dyson, A. E. (1971). (eds.). *The Black Papers on Education* (London, Davis-Poynter).

Cresswell, M. J. (1987). Describing Examination Performance: grade criteria in public examinations. *Educational Studies*, **13**(3), 247–65.

Cresswell, M. J. (1990). Gender Effects in GCSE—Some Initial Analyses (Paper prepared for a Nuffield Seminar at University of London Institute of Education on 29 June 1990) (Unpublished Associated Examining Board Research Report—RAC/517).

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*, **1**(1), 45–61.

Cresswell, M. J. (1995). Technical and Educational Implications of using Public Examinations for Selection to Higher Education. In T. Kellaghan (ed.), *Admission to Higher Education: Issues and Practice* (Dublin, Educational Research Centre and Princeton, International Association for Educational Assessment).

Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum Embedded Examinations: Judgemental and Statistical Approaches. In H. Goldstein and T. Lewis (eds.). *Assessment: Problems, Developments and Statistical Issues* (London, Wiley).

Cresswell, M. J. (1997a). *Examining Judgements: Theory and Practice of Awarding Public Examination Grades* (PhD thesis, University of London Institute of Education).

Cresswell, M. J. (1997b). Can Examination Grade Awarding be Objective and Fair at the Same Time? Another Shot at the Notion of Objective Standards (Unpublished Associated Examining Board Research Report—RAC/733).

Cresswell, M. J. and Houston, J. G. (1991). Assessment of the National Curriculum—some fundamental considerations. *Educational Review.* **43**, 63–78.

Cressy, D. (1980). *Literacy and the Social Order: reading and writing in Tudor and Stuart England* (Cambridge, Cambridge University Press).

Damasio, A. R. (1995). *Descartes Error: Emotion, Reason and the Human Brain* (London, Papermac).

Davis, E. (1993). *Schools and the State* (London, Social Market Foundation).

Dean, C. (1998). Standards are not parents' top priority. *Times Educational Supplement*, 9 October.

Dearing, R. (1995). *Review of the 16–19 qualifications* (London, Department of Education).

Dennett, D. (1993). *Consciousness Explained* (London, Penguin).

Department for Education and Employment (DfEE). (1997). *Excellence in Schools* (London, Stationery Office).

Department of Education and Science (1967). *Children and Their Primary Schools. A Report of the Central Advisory Council for Education (England).* ii (London, DES).

Devine, M., Hall, J., Mapp, J. and Musselbrook, K. (1996). *Maintaining Standards: Performance at Higher Grade in Biology, English, Geography and Mathematics* (Edinburgh, Scottish Council for Research in Education).

Devlin, K. (1997). *Goodbye Descartes: The End of Logic and the Search for a New Cosmology of the Mind* (New York, Wiley).

Dore, R. (1996). *The Diploma Disease.* 2nd edition (London, Institute of Education).

Dreyfus, H. L. (1992). *What computers still can't do: a critique of artificial reason.* (Cambridge Mass., MIT Press).

Eagleton, T. (1993). *Literary Theory: An Introduction* (Oxford, Blackwell).

Eiser, J. R. (1990). *Social Judgement* (Milton Keynes, Open University Press).

Elwood, J. and Comber, C. (1996). *Gender differences in examinations at 18+* (London, Institute of Education).

Firestone, W. A. (1998). A Tale of Two Tests: Tensions in Assessment Policy. *Assessment in Education*, **5**(2), 175–192.

Fletcher, S. (1980). *Feminists and Bureaucrats. A study in the development of girls' education in the nineteenth century* (Cambridge, Cambridge University Press).

Fogelin, R. J. (1967). *Evidence and Meaning: Studies in Analytic Philosophy* (London, Routledge).

Forrest, G. M. and Orr, L (1984). *Grade Characteristics in English and Physics* (Manchester;,Joint Matriculation Board).

Foxman, D., Ruddock, G. and McCallum, I. (1990). *APU mathematics monitoring 1984–88 (Phase 2)* (London, Schools Examination and Assessment Council).

Fremer, J. (1989). Testing Companies, Trends and Policy Issues: A current view from the testing industry. In B. R. Gifford (ed.), *Test Policy and the Politics of Opportunity Allocation: The Workplace and the Law* (Boston, Kluwer).

French, S., Slater, J. B., Vassiloglou, M. and Willmott, A. S. (1987). *Descriptive and Normative Techniques in Examination Assessment* (Oxford, UODLE).

Galton, M. (1998). Back to consulting the ORACLE. *Times Educational Supplement*, 3 July.

Gierl, M. J. and Rogers, W. J. (1996). Factor analysis of the Test Anxiety Inventory using Canadian high school students. *Educational and Psychological Measurement*, **56**, 315–324.

Goldstein, H. (1983). Measuring Changes in Educational Attainment Over Time: Problems and Possibilities. *Journal of Educational Measurement*, **20**, 369–78.

Goldstein, H. (1995). *Interpreting International Comparisons of Student Achievement* (Paris, UNESCO).

Goldstein, H. (1996a) (ed.). *Assessment in Education, 3,* 2. Special Issue: The IEA Studies.

Goldstein, H. (1996b). International Comparisons of Student Achievement. In Little and Wolf (1996).

Goldstein, H. (1999). Performance Indicators in Education. In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold).

Goldstein, H. and Cresswell, M. J. (1996). The comparability of different subjects in public examinations: a theoretical and practical critique. *Oxford Review of Education*, **22**(4), 435–42.

Goldstein, H. and Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology*, **42**, 139–167.

Good, F. J. and Cresswell, M. J. (1988a). *Grading the GCSE* (London, Secondary Examinations Council).

Good, F. J. and Cresswell, M. J. (1988b). *Differentiated Assessment: Grading and Related Issues* (London, Secondary Examinations Council).

Gould, S.J. (1984). *The Mismeasure of Man* (London, Penguin).

Gray, J., Hopkins, D., Reynolds, D., Wilcox, B., Farrell, S. and Jesson, D. (1999). *Improving Schools: Performance and Potential* (Milton Keynes, Open University Press).

Gray, J., McPherson, A. and Raffe, D. (1983). *Reconstructions of Secondary Education* (London, Routledge).

Green, A., Leney, T. and Wolf, A. (1997). *Convergences and Divergences in European Education and Training Systems* (Brussels, EC Directorate-General XXII (Education, Training and Youth)).

Green, A., Wolf, A. and Leney, T. (1999). *Convergence and Divergence in European Education and Training Systems* (London, Institute of Education).

Hacking, I. (1965). *The Logic of Statistical Inference* (Cambridge, Cambridge University Press).

Hacking, I. (1990). *The Taming of Chance* (Cambridge, Cambridge University Press).

Hambleton, R. K. and Zaal, J. N. (eds.) (1991). *Advances in Educational and Psychological Testing* (Boston, Kluwer).

Hargreaves, D. H. (1996). Teaching as a research-based profession: policies and prospects (Teacher Training Agency annual lecture).

Heath, A. F. and Clifford, P. (1990). Class inequalities in education in the twentieth century. *Journal of the Royal Statistical Society*, series A, **153**, 1–16. .

Holland, P. W. and Rubin, D. B. (1982). *Test Equating* (New York, Academic Press).

Hollis, M. and Lukes, S. (1982). (eds). *Rationality and Relativism* (Oxford, Blackwell).

Holmes, E. (1911). *What Is and What Might Be* (London, Constable).

Jencks, C. (1972). *Inequality: A Reassessment of the Effect of Family and Schooling in America* (New York, Basic Books).

Johnson, V. E. (1997). An alternative to the traditional GPA for evaluating student performances. *Statistical Science*, **12**, 251–278.

Kelly, A. (1976). A study of the comparability of external examinations in different subjects. *Research in Education*, **16**, 37–63.

Kilpatrick, J. and Johansson, B. (1994). Standardised Mathematics Testing in Sweden: The Legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, *6–30.*

Koretz, D., Broadfoot, P. and Wolf, A. (1998) (eds.). *Assessment in Education*, **5**(3) (Special Issue on Portfolios and Records of Achievement).

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, second edition (Chicago, University of Chicago Press).

Lakatos, I. (1974). *Proofs and Refutations: the Logic of Mathematical Discovery* (Cambridge, Cambridge University Press).

Little, A. (1996) (ed.). *Assessment in Education*, *4*(1) (Special Issue: The Diploma Disease Twenty Years On).

Little, A., Wang Gang, and Wolf, A. (1995) (eds.). *Sino-British Perspectives on Educational Assessment* (London, ICRA, Institute of Education).

Little, A. and Wolf, A. (1996) (eds.). *Assessment in Transition: Learning, monitoring and selection in international perspective* (Oxford, Pergamon).

Long, H. A. (1985). Experience of the Scottish Examinations Board in developing a grade-related criteria system of awards (Paper presented at the 11th annual conference of the International Association for Educational Assessment held in Oxford, England).

Macaulay, Lord (1898). *Collected Works*, 12 vols. (London, Longmans Green).

Mackenzie, D. A. (1981). *Statistics in Britain 1865–1930. The Social Construction of Scientific Knowledge* (Edinburgh, Edinburgh University Press).

McKenzie, D. (1994). The irony of educational review. *New Zealand Annual Review of Education*, **4**, 247–59.

McLean, L. D. (1996). Large-Scale Assessment Programmes in Different Countries

and International Comparisons. In H. Goldstein and T. Lewis (eds.), *Assessment: Problems, Developments and Statistical Issues* (Chichester, Wiley).

McPherson, A. and Willms, J. D. (1987). Equalisation and improvement: some effects of comprehensive reorganisation in Scotland. *Sociology*, **21**, 509–39.

Madaus, G. and Raczek, A. (1996). Turning Point for Assessment: Reform Movements in the United States. In Little and Wolf (1996).

Menet, J. (1874). *A Letter to a Friend on the Standards of the New Code of the Education Department* (London, Rivingtons).

Morrison, H. G., Busch, J. C. and D'arcy, J. (1994). Setting reliable national curriculum standards: a guide to the Angoff procedure. *Assessment in Education*, **1**, 181–199.

Murphy, R. J. L. (1982). Sex differences in Objective Test performance. *British Journal of Educational Psychology*, **52**, 213–19.

Murphy, R. J. L., Burke, P., Cotton, T., Hancock, J., Partington, J., Robinson, C., Tolley, H., Wilmut, J. and Gower, R. (1996). *The Dynamics of GCSE Awarding: Report of a project conducted for the School Curriculum and Assessment Authority* (London, SCAA).

Newcastle Report (1861). *Report of the Commissioners appointed to inquire into the State of Popular Education in England*, PP 1861 XXI (ii) (London).

Newton, P. (1996). The reliability of marking of GCSE scripts: Mathematics and English. *British Educational Research Journal*, **22**, 405–20.

Newton, P. (1997a). Measuring comparability of standards between subjects: why our statistical techniques do not make the grade. *British Educational Research Journal*, **23**(4), 433–49.

Newton, P. (1997b). Examining Standards Over Time *Research Papers in Education*, **12**(3), 227–48.

Orr, L. and Forrest, G. M. (1984). *Investigation into the relationship between grades and assessment objectives in History and English examinations* (Manchester, Joint Matriculation Board).

Orr, L. and Nuttall, D. L. (1983). *Determining Standards in the Proposed Single System of Examinations at 16+* (London, Schools Council).

Paterson, L. (1992). The influence of opportunity on aspirations among prospective university entrants from Scottish schools, 1970–1988. *Statistics in Society, Journal of the Royal Statistical Society*, series A, **155**, 37–60.

Paterson, L. (1995). Social origins of under-achievement among school-leavers. In L. Dawtrey, J. Holland, M. Hammer and S. Sheldon (eds.), *Equality and Inequality in Education Policy* (Milton Keynes, Open University Press).

Paterson, L. (1997). Student achievement and educational change in Scotland, 1980–1995. *Scottish Educational Review*, **29**, 10–19.

Paterson, L. (1998). The Scottish parliament and Scottish civil society: which side will education be on? *Political Quarterly*, **69**, 224–33.

Paterson, L. (forthcoming). Scottish traditions in education. In H. Holmes (ed.), *Compendium of Scottish Ethnology, vol. 11* (Edinburgh, Scottish Ethnological Research Centre).

Paterson, L. and Raffe, D. (1995). Staying on in full-time education in Scotland. *Oxford Review of Education*, **21**, 3–23.

Payne, J. (1872). 'Why are the Results of our Primary Instruction so Unsatisfac-

tory?', *Transactions of the National Association for the Promotion of Social Science.*

Phillips, M. (1996). *All Must Have Prizes* (London, Little, Brown and Company).

Pirsig, R. M. (1974). *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (London, Bodley Head).

Plewis, I. (1998). Inequalities, Targets and Zones. *New Economy*, **5**, 104–8.

Plewis, I. (1999). What's Worth Comparing in Education? In D. Dorling and S. Simpson (eds.). *Statistics in Society* (London, Arnold), 273–80.

Pole, D. (1961). *Conditions of Rational Inquiry: A Study in the Philosophy of Value* (London, Athlone).

Power, M. (1997). *The Audit Society: Rituals of Verification* (Oxford, Oxford University Press).

QCA (1998). *GCSE and GCE A/AS code of practice* (London, Qualifications and Curriculum Authority).

Reynolds, D., Creemers, B. P. M., Stringfield, S. and Teddlie, C. (1998). Climbing an educational mountain: conducting the International School Effectiveness Research Project. In G. Walford, *Doing research about education* (Lewes, Falmer Press).

Roach, J. P. C. (1971). *Public Examinations in England 1850–1900* (Cambridge, Cambridge University Press).

Robertson, C. (1992). Routes to higher education in Scotland. *Scottish Educational Review*, **24**: 3–16.

Rose, S. (1997). *Lifelines, Biology, Freedom, Determinism* (London, Penguin).

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education*, **13**, 191–209.

Schools Council (1979). *Standards in Public Examinations: Problems and Possibilities*, Report from the Schools Council Forum on Comparability (London, Schools Council).

SEC (1984). *The development of Grade-related Criteria for the General Certificate of Secondary Education—a briefing paper for working parties* (London, Secondary Examinations Council).

SEC (1985). *Reports of the Grade-related Criteria Working Parties* (London, Secondary Examinations Council).

SEC (1986). Draft Grade Criteria. *SEC News Number 2* (London, Secondary Examinations Council).

SEC (1987). Grade Criteria—Progress Report. *SEC News Number 6* (London, Secondary Examinations Council).

Shavit, Y. and Blossfeld, H. P. (1993). *Persistent Inequality: Changing Educational Attainment in Thirteen Countries* (Boulder, Col., Westview Press).

Skolöverstyrelsen (1980). Quoted in J. Kilpatrick and B. Johansson (1994). Standardised Mathematics Testing in Sweden: The legacy of Frits Wigforss. *Nordic Studies in Mathematics Education*, **1**, 6–30. .

Smith, J. V. and Hamilton, D. (1980) (eds). *The Meritocratic Intellect* (Aberdeen, Aberdeen University Press).

Start, B. and Wells, K. (1972). *The trend of reading standards* (Slough, National Foundation for Educational Research).

Stedman, L. C. (1998). An Assessment of the Contemporary Debate over US

Achievement. In D. Ravitch (ed.), *Brookings Papers on Education Policy* (Washington DC, Brookings Institution Press), 53–119.

Stephens, W. B. (1987). *Education, Literacy and Society, 1830–70: the geography of diversity in provincial England* (Manchester, Manchester University Press).

Sutherland, G. (1973a). *Policy-Making in Elementary Education 1870–1895* (Oxford, Clarendon Press).

Sutherland, G. (1973b) (ed.). *Matthew Arnold on Education* (London, Penguin).

Sutherland, G. (1984). *Ability, Merit and Measurement. Mental testing and English education 1880–1940* (Oxford, Clarendon Press).

The Scotsman Education (1998). 30 September: 4–5.

Thom, D. (1986). The 1944 Education Act: the 'art of the possible. In Harold L. Smith (ed.), *War and Social Change: British Society in the Second World War* (Manchester, Manchester University Press), 101–28.

Vincent, D. (1989). *Literacy and Popular Culture: England 1750–1914* (Cambridge, Cambridge University Press).

Walden, G. (1996). *We Should Know Better: solving the educational crisis* (London, Fourth Estate).

Wang Binhua (1995). Comparing HSCE in the People's Republic of China and GCSE in England. In Little, Wolf and Wang Gang (1995).

Wang Gang (1995). The Development of Public Educational Examinations in China from 1980. in Little, Wolf and Wang Gang (1995).

Wiliam, D. (1996a). Meanings and Consequences in Standard Setting. *Assessment in Education*, **3**(3), 287–307.

Wiliam, D. (1996b). Standards in examinations: a matter of trust? *The Curriculum Journal*, **7**(3), 293–306.

Wilmut, J. and Rose, J. (1989). *The Modular TVEI Scheme in Somerset: its concept, delivery and administration* (Report to the Training Agency of the Department of Employment, London).

Wolf, A. (1995). Competence Based Assessment (Buckingham, Open University Press).

Wolf, A. and Steedman, H. (1998). Basic Competence in Mathematics: Swedish and English 16 year olds. *Comparative Education*, **34**, 3.

Wood, R. (1991). *Assessment and Testing: A survey of research* (Cambridge, Cambridge University Press).

Young, M. (1958). *The Rise of the Meritocracy 1870–2033* (London, Penguin).