Comparative Education
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cced20

International Comparisons of School Effectiveness: The second component of the 'crisis account' in England?

Stephen Gorard

Available online: 28 Jun 2010

To cite this article: Stephen Gorard (2001): International Comparisons of School Effectiveness: The second component of the 'crisis account' in England?, Comparative Education, 37:3, 279-296

To link to this article: http://dx.doi.org/10.1080/03050060120067785

Comparative Education *Volume 37 No. 3 2001 pp. 279–296*

International Comparisons of School Effectiveness: the second component of the 'crisis account' in England?

STEPHEN GORARD

ABSTRACT *Schools and students in Britain have been compared unfavourably with those elsewhere for many years now. International comparisons of school effectiveness and outcomes have been used to suggest that British schools are underperforming, and consequently that British students are underachieving. This has led to repeated calls by researchers and politicians for policy borrowing from countries with more 'successful' educational systems. In the same way, the growth of 'home–international' comparisons has suggested marked differences between the increasingly devolved regions within Britain. This view of the relative ineffectiveness of schools has permeated both official publications and regional attainment targets in the 1990s. This article considers some of these findings from the perspective of England, and the associated problems in studies involving international comparisons. It concludes that, viewed proportionately and in the light of background factors, there is little evidence of either failing or falling standards. Although, as always, absence of evidence should not be taken as evidence of absence there is certainly little reason to agree with the 'crisis account' of English schooling. The implications for the role of international league tables are discussed briefly. Cross-national comparisons, seen in this light, are open to misinterpretation and are therefore potentially dangerous.*

'A Plague on all our Schools'

This article briefly considers the role of international comparisons of school effectiveness in educational policy formation. It shows that the results need very careful contextualisation, otherwise they are open to misinterpretation. This is done by re-examining an apparent crisis in British schools, with particular reference to schools in England (the largest of the four home countries in the UK). However, the findings and their implications have wider relevance for any country faced by apparently unflattering references within international 'league tables'.

The view that schools in England (and Wales) are underperforming, even failing, in international terms forms an important part of what I have termed the 'crisis account' of UK schooling. This crisis account is a shared perspective of a loose alliance of researchers and other commentators who apparently recall some golden age of schooling, when educational standards were generally higher, and social justice was greater. Since that time, divisions

Correspondence to: Dr Stephen Gorard, Cardiff University School of Social Sciences, Glamorgan Building, King Edward VII Avenue, Cardiff CF10 3WT, UK. E-mail: gorard@cardiff.ac.uk

ISSN 0305-0068 print; ISSN 1360-0486 online/01/030279-18 © 2001 Taylor & Francis Ltd
DOI: 10.1080/03050060120067785

within UK schools and their underachievement in international terms are supposed to have increased.

If the first component of this alleged crisis is the view that educational opportunities are increasingly polarised (see Gorard, 2000a), then a belief in poor overall educational standards forms the second component. While there may be differences in the motivation of the crisis disciples pushing each component, the overall effect of all these writers is distressing—distressing and wrong. I have written about the first component of the alleged crisis in detail elsewhere (Gorard, 2000b). This article starts by outlining some of the claims about the relative failure of schools in England and Wales, and some of the forms of evidence used to support these claims. It continues with a more detailed consideration of the difficulties of assessing differences in educational standards over time and space, and therefore of making sensible comparisons between international and even regional school systems.

Failing and Falling Standards?

Standards in British education in 1969 were described by Cox & Dyson (1990) as lower than in 1929, indeed lower even than in 1914. This decline was attributed by them to, among other things, a philosophy of egalitarianism rather than equality of opportunity, to the comprehensivation of schools rather than selection, and to the use of 'progressive' rather than traditional teaching methods. Thus, as early as 1975 a future Minister of Education wrote that, 'British children are the most ill-mannered, undisciplined and selfish in the world' (Boyson, 1975, p. 119), and by 1996 an influential writer on education believed that, 'there is now a yawning gap between the standards reached by British schoolchildren and their counterparts in Europe or Japan' (Phillips, 1996, p. 3). This view finds an ally in Barber (1996), who claims that there is indeed a crisis in education today as it faces the twin threats of global competition and local moral decline, manifested in a growing gap between the best and worst performing students. George Mudie (while Minister for Lifelong Learning) recently claimed that in any other industry a performance level like that of British education 'would result in the companies concerned going out of business' (Skills and Enterprise Network, 1999, p. 1). According to a personnel director of Siemens Electronics, the workforce in Britain is crucially short of key skills.

> The fact that we rate after Greece and New Zealand in world terms in literacy and numeracy was deeply worrying … Llew Aviss finished with an example of how far behind we were in maths from one of his former employers, Fujitsu. It was a key issue for Japanese colleagues who were sending their children to school in Britain. They would query why their children were so much further behind where they would be in Japan … and this was only at the primary school level. (Skills and Enterprise Network 1999, p. 5)

There is plenty of practical and academic research evidence that can be presented in support of these views. The Office for Standards in Education (OFSTED) and the National Foundation for Educational Research (NFER) have confirmed that there has been a deterioration of standards in several areas of education, while a previous Chief Inspector of Schools believes that, 'where we fail badly, in comparison with other developed countries, is with the broad range of ordinary pupils' (Bolton, 1993, p. 12). The Second International Maths Study (SIMS) revealed not only that students in England and Wales were achieving relatively low marks in international comparisons, but also that standards in England and Wales had actually

declined between 1964 and 1981 (Cresswell & Gubb, 1990). By 1981, it was estimated that the mathematical attainment of English 13-year-old students was four years behind their peers in Japan (Prais, 1990). In the Third International Maths and Science Study (TIMSS), the mean worldwide score for attainment in mathematics among all Year 9 students was 513, whereas for England the score was 506, and this was for children who had received, on average, an extra year of formal schooling in comparison to the sample from other countries (Beaton *et al.*, 1996). By comparison, the equivalent score in Ireland was 527, in France 538, in Japan 605 and in Singapore a massive 643. A recent Department for Education and Employment (DfEE) report stated that, 'Britain performs badly in basic skills stakes' (Skills and Enterprise Briefing, 1999), showing that of 12 OECD countries Britain was second only to Poland and Ireland in the proportion of adults with low levels of literacy and numeracy skills. The DfEE therefore claim that the scale of the problem of literacy and numeracy in England is 'enormous', and cite an example in which 20% of adults could not locate the page reference for 'plumbers' in an alphabetical index (DfEE, 1999).

One reason for this overall depressing state of affairs may be that teachers now spend less time teaching than in previous generations, and more time in bureaucratic tasks (Barber, 1996). Porter (1990) blames both students and teachers for the poor competitive position of British industry, and refers to the educational system as 'a major barrier to … sustaining competitive advantage in industry' (p. 7). 'British children are taught by teachers less qualified than those in many nations, receive less training in maths and science, put in fewer hours, and drop out more' (p. 7). In an influential summary of British education at the end of the 1990s, Bentley (1998) describes the full malaise of limited literacy skills among students, large numbers of school leavers with no qualifications, an increasing performance gap between girls and boys, and growing indicators for truancy and exclusion from school. Even where examination outcomes have improved over time, as evidenced by an increasing number of A grades or the introduction of the A* grade at GCSE, this can be attributed by 'crisis' commentators to a simple lowering of standards (or 'dumbing down' as it has become popularly known). For example, the mathematical ability of A-level candidates and entrants to higher education has been reported as being in decline over time, even where candidates are matched in terms of their A-level grades (Kitchen, 1999), thus showing that equivalent-sounding qualifications are now worth less than in previous years, not only in exchange value but also in actual fact.

Chris Woodhead, the former Chief Inspector for Schools in England, was also quoted recently as saying that, 'a staggering 55% of boys have not learnt to write properly … These are the 11-year-olds who will grow into the 16-year-olds who leave schools unemployed and … unemployable' (Woodhead, 1999). A recent MORI poll on family life in Britain found that the public perception of Britain's education system was poor compared to elsewhere in the European Union, and a majority believed that standards in schools are falling (Petre, 1999). The same report stated that universities are now giving degree students lessons in literacy, and that the demand for such courses is increasing. This is seen by critics as evidence that schools are failing to teach basic skills even to the most talented individuals in society.

Despite ever-increasing government expenditure on education, longer and longer initial schooling, and lower pupil:teacher ratios, standards of literacy among school-leavers have not improved and may even have fallen since 1944 according to some influential observers (e.g. Boyson, 1975). Stories about poor educational performance in Britain, with its long tail of underachievement, are therefore not new, and the evidence has often been presented before (e.g. Postlethwaite, 1985; IEA, 1988). Reading standards among seven and eight year olds were found to have dropped in the 1980s, just as the number skills among 11 to 15 year olds

were found to have deteriorated (National Commission on Education, 1993). However, this crisis appears to have peaked in the 1990s. In a speech at the Lord Mayor's banquet at the Guildhall in London in 1993, the then Prime Minister, John Major, said: 'Let me give you just one example. In arithmetic, 13-year-olds were asked to multiply 9.2 by 2.5. In Korea and Taiwan 70% got it right, in Western Europe 55%, in England 13%.' According to the Director General of the CBI, the UK was recently ranked twenty-second out of 22 in an international comparison on the ability of its education system to meet the needs of international competition in the 1990s (Corbett, 1993).

A survey of 500 head teachers found that standards of literacy (among primary school-leavers) were still falling in 1991 (Phillips, 1996), while 'a core of about 20% of 7 year-olds have failed to reach the standard for their age since the first of these tests ... was held in 1991 ... [and] the first tests for 11-year-olds held in 1995 revealed that more than half were not up to scratch' (p. 4). Among school leavers 'the UK ... lags behind many of its European and international competitors' (ETAG, 1998, p. 11). For example, at age 16 the percentage gaining a GCSE grade C + (or equivalent) in mathematics, a science, and the national language was 27% in England in 1991. In Japan the corresponding figure was 50%, in Germany 62%, and in France 66% (Phillips, 1996). By the age of 18, a comparable qualification was attained by 29% in England, 48% in France, 68% in Germany and 80% in Japan (National Commission on Education, 1993). In the population as a whole, only 45% of those in the UK have attained the equivalent of a National Vocational Qualification (NVQ) Level 2, whereas in France the figure is 65% and in Germany 70% (ETAG, 1998).

A series of international comparisons over the last 20 years, some based on examination outcomes and some on purpose-designed tests, have confirmed these relatively poor figures from British (primarily English) students. The comparisons tend to concentrate on the subjects of mathematics and science, in which it is seen as easier to overcome the difficulties of 'common curriculum currency' since of all school subjects these have the fewest cultural influences, according to Reynolds & Farrell (1996) and others. In a review of these international comparisons, Reynolds & Farrell claim that the use of standardised attainment tests in all participating countries solves the 'common currency' problem. Of course, any differences could have non-educational reasons and the difficulty for any research in this area lies chiefly in separating out the impact of educational and socio-economic determinants of academic performance (see below). Nevertheless, Reynolds & Farrell claim that as there are no *known* non-educational causes of differential attainment in mathematics and science, 'it is clear, then, that the educational systems of different societies are key factors in determining their educational attainment' (1996, p. 52). It is therefore noteworthy that in all the studies reviewed the attainment of students from England (and Wales in some studies) was poor overall. In addition this attainment has been getting worse, relative to other countries, since the 1960s. The student performance in England is generally of greater variability than in most countries, with a few very good scores and a lot of very poor ones. This is despite a longer than average experience of compulsory schooling, and a low response rate from England which is liable to lead commentators to over-estimate the scores. 'It would in our view need rather more than the above list of caveats to persuade one that the English performance is anything other than poor' argue Reynolds & Farrell (1996, p. 53). It is on the basis of this conclusion that Reynolds & Farrell go on to examine potential reasons for this poor national performance, and to suggest models for improvement based on the classroom practice of more 'successful' countries from the Pacific Rim as evidenced by the TIMSS. The report gained wide publicity and media attention, and has led to significant changes of policy and practice especially for trainee teachers.

How Do We Judge Falling Standards?

The issue of judging standards over time is a difficult one to investigate without having a close definition of the term 'standard'. As an illustration of how elastic the term can be, consider the very real situation in which an educational attainment indicator such as A-level performance becomes more common over a period of ten years. One group of politicians may claim that standards have therefore improved demonstrably as more students now attain the A-level standard. Their opponents may claim that standards have fallen, since the A-level is now demonstrably easier to obtain and also worth less in exchange. A similar example was outlined above where the mathematical ability of university students, who had achieved the same grade at A-level, was seen to have declined (Kitchen, 1999). However, on closer inspection the changes could be interpreted as related to concurrent changes in the policy of university recruitment, and changes in the relative popularity of different courses.

The point to be made here is that knowledge is not a static commodity, and therefore that comparisons of changes over time in school attainment have to try and take these changes into account. One analogy for the complaint by the National Commission on Education (1993) that number skills amongst 11–15 year olds had deteriorated would be the clear drop over the last millennium in archery standards among the general population. Nuttall (1979) used the example of the word 'mannequin' to make the same point. If the number of children knowing the meaning of this word drops from the 1950s to the 1970s, is this evidence of some kind of decline in schooling? Perhaps it is simply evidence that words and number skills have changed in their everyday relevance. On the other hand, if the items in any test are changed to reflect these changes in society, then how do we know that the test is of the same level of difficulty as its predecessor? In public examinations, by and large, we have until now relied on norm-referencing. That is, two tests are declared equivalent in difficulty if the same proportion of matched candidates obtain each graded result on both tests. The assumption is made that the actual standards of each annual cohort are equivalent, and it is these that are used to benchmark the assessment. How, then, can we measure changes in standards over time? If the test is not norm-referenced, how can we tell that apparent changes over time are not simply evidence of differentially demanding tests? This apparently insuperable problem has, to my mind, not been adequately addressed by 'crisis' commentators.

It has been claimed that the level of attainment required to gain Level 4 at KS2 has fallen over time. The evidence cited for this assertion is that whereas students needed 52% to gain Level 4 English in 1997, the corresponding figures for 1998 and 1999 were 51% and 47% (Cassidy, 1999a). The response from the English Qualifications and Curriculum Authority (QCA) is that percentages are bound to change over time as the difficulty of the tests varies year-on-year, but that these differences are not educationally significant. A counter response has been that the QCA deliberately reduced the threshold because David Blunkett (Secretary of State for Education) had staked his career on 80% of 11 year olds gaining Level 4 by 2002. Since in 1998 only 65% of the population gained Level 4, it is claimed that while the target has been retained the pass mark has been conveniently lowered. An independent enquiry was ordered, the results of which have mainly supported the QCA position. This debate encapsulates the problems of discussing changes in assessments over time (also see below).

When serious attempts have been made to compare standards of attainment over time, and taking into account all of the above caveats, the results are generally that standards are *not* falling. In some cases there is no firm evidence of change, and in others there are improvements over time. For example, an analysis of successive GCSE cohorts from 1994 to 1996 found a significant improvement in performance over time (Schagen & Morrison, 1998). It is possible to question the reality of this improvement in strict criterion-referenced

terms, but there is at any rate no evidence of any decline, and some suggestion that things are actually getting better.

The Problems of Assessment

This brief section indicates a few of the common problems faced by researchers when judging the reliability and validity of formal school assessments. Since the majority of publicly available (rather than anecdotal) evidence cited for the crisis account of British schooling is of this form, it is important to look, even if only briefly, behind the facade of objective rigorous testing procedures and to consider the possibility that any sets of figures will be based at least in part on subjective judgements, or error components, or worse. For the purpose of this article, the following question has to be asked—if assessments are not totally reliable, then how can one attribute the differences between countries to real educational differences? Put simply, the size of the differences between countries would have to be larger than the variation in the results for each country which is attributable to these errors for them to be even considered as potential evidence of real educational differences.

Britain may be unique among the countries shown in Table I in using different regional authorities (local examination boards) to examine what are meant to be national assessments at 16 + and 18 + (Noah & Eckstein, 1992). This raises an issue of whether the same qualification is equivalent between these boards in terms of difficulty. It is already clear that even qualifications with the same name (e.g. GCSE History) are not equivalent in terms of subject content as each board sets its own syllabus. Nor are they equivalent in the form of assessment, or the weighting between components such as coursework and multiple choice. Nor is there any evidence that the different subjects added together to form aggregate benchmarks are equivalent in difficulty to each other; yet the standard GCSE benchmark gives the same value to an A* in Music, a B in Physics, and a C grade in Sociology. Nor is there evidence that qualifications with the same name are equally difficult from year to year. In fact, comparability is an issue between boards in any subject, between different years in the same subject/board, between the subjects in one board, and even between the alternative syllabuses in any board and subject. All of these are very difficult to determine, especially as exams are neither accurate nor particularly reliable in what they measure (Nuttall, 1979). Pencil-and-paper tests have little generalisable validity, and their link to other measures such as occupational competence is also generally very small (Nuttall, 1987).

Thus, it is not surprising that even the supposedly national system of statutory assessment is producing a flood of complaints about irregularities and inconsistencies. There is evidence, for example, that class teachers, and sometimes even head teachers, reveal the content of national tests to their students before they sit the tests. In one school two classes took a SAT test simultaneously. One assessment took place in one and a half hours of strict silence. The other assessment took all morning, and the teacher was heard to give clues by standing over a child and saying 'you may need to get your rubber out for that one'. How can the ensuing results be considered comparable, even though they are from the same school?

Some heads condemned the marking system for national tests after it was claimed that scripts had been lost and tests scores added up incorrectly. One school checked the results after they were returned to them and found nine errors in adding up the marks in 60 KS3 mathematics scripts (Cassidy, 1999c). According to some English teachers, students taking that year's tests in English for 13/14 year olds would have needed a reading age of 16 (Cassidy, 1999b). Of course, claims such as these may be no more valid than the evidence in support of a crisis, but the fact that there are even possibilities of this kind does help to lift

some of the facade of rigorous reliability in examination processes. It is clear that even the narrow version of propositional knowledge tested by examinations is very difficult to assess, not least because there is general confusion between the use of examinations for formative, summative and target purposes (Holt, 1981; Daugherty, 1995). In a letter to *The Times* in 1976, Nuttall wrote 'The message is clear: examination standards do not necessarily tell us anything about educational standards'.

What Difference Does a School Make?

Much recent educational research stems from the influence of the school effectiveness movement. This field of study has attempted to describe the characteristics of a successful school in a way that could form the basis of a blueprint for school improvement. Ironically, the major undisputed outcome of all of this work has been the reinforcement of the importance of the non-school context. National systems, school sectors, schools, departments and teachers combined have been found to explain approximately zero to 20% of the total variance in school outcomes (depending upon the study). The remainder of the variance in outcomes is explained by student background, prior attainment and error components. Despite this, most educational policies are based upon comparisons between schools that do not take these incontrovertible findings into account. Such policies include league tables of results, programmes of inspection, and national and regional targets, all of which have presented attainments in raw-score forms. Surprisingly, several of the commentators who have been central to the school effectiveness movement, proposing models of value-added analysis, are also those proposing regional and international comparisons based on raw-score differences.

The very large-scale studies by Coleman *et al.* (1966) and Jencks *et al.* (1972) showed that once individual student characteristics had been taken into account, very little of the variance in school examination outcomes was left to be explained by a 'school effect'. In essence, what these studies showed is that specific schools as institutions, and schooling as a process, may have very little impact on the results attained by students. To put it simply, any individual would be expected to attain pretty much the same results whichever school they attended. The variability in student outcomes can be almost entirely explained by student 'context' factors such as family background. More recently, the school effectiveness movement (SEM) claims to have uncovered systematic variation between schools as institutions, and has attributed these large differences (in their terms) to school effects. Put simply, they are saying in contradiction to the earlier studies that it does matter which school a student attends, and that there are therefore 'good' schools and 'bad' schools.

To some extent the differences between these two groups of researchers can be seen as rhetorical rather than real. For the SEM group, Reynolds (1990) states that up to 8% of the variance in school results is due to school effects rather than individual characteristics, and he calls these 'large school effects' (p. 154). The OECD (1995), on the other hand, says that most of the variation in school results can be explained by school input factors (such as student socio-economic indicators or scores of prior attainment). According to their review, the residual that could be explained by schools is *only* 12%. Rather like the story of the full/empty beer glass, 8% is a large amount to one school effectiveness researcher, but 12% is only 12% to another. Harker & Nash (1996) claim that the school effect on outcomes in New Zealand is 'small', but then show that only between two-thirds and three-quarters of variance in the School Certificate is due to the ability and social mix of each school. From the United States, the evidence since the early studies by Coleman *et al.* (1966) and Jencks

et al. (1972) has been largely pessimistic about school effects. In all studies the effect is very small, and the larger the sample used, the weaker is the evidence of any school effect at all (Shipman, 1997). The social background of the children is still apparently all-important.

The proportion that is, or could be, explained by school-level processes (the difference a school actually makes) is fairly well agreed within limits. Daly (1991) estimates between 8–10%, Creemers (1994) attributes 12–18% of the variance to schools once the background of students is taken into account, while Stoll & Fink (1996) put the figures at 8–14%. In a review of effectiveness studies, Gray & Wilcox (1995) attribute from 2–10% of the variance to school effects. Lauder *et al.* (1999) demonstrate that 80–99% of the variance in school outcomes is attributable to measures in relation to the individual student (actual figure depending on the specific outcome measure). If these individual-level variables (such as socio-economic status (SES), ethnicity, gender and prior attainment) are aggregated to a school level there may also be a 'school-mix' effect. The individual student characteristics, the mix, and related non-academic items like school size and the stability of the number on roll, leave very little variance to be explained by both school educational processes and error components combined. The specific figures vary from study to study, but the overall pattern is clear. Much recent work shows that the adjusted differences between schools are very small, so it is difficult to distinguish schools from each other (Ouston, 1998), especially as 'outlier' scores such as those from children with special needs are routinely eliminated before analysis (Hamilton, 1998).

For some writers, school effectiveness has become a kind of cult, and therefore difficult to argue against. Several of its most devoted adherents have become government advisers, and so the movement itself is becoming more and more part of the official discourse, mixed with an economic vocabulary about targets for lifelong learning, and market-driven performance indicators. One of these advisers in the UK is Barber (1996), who concludes from a review of the evidence that:

> The research into school effectiveness over the last two decades has made an immense contribution to our understanding of school performance. Whereas in the 1960s and early 1970s the prevailing view of education and social researchers was that the effect of school on a pupil's performance was negligible in comparison to the impact of social class and upbringing, it is now demonstrable that schools make a significant difference to how well children do. (p. 127)

When researchers have attempted to relate this small school-effect to school characteristics and processes, so producing a blueprint for school improvement, the results often have generally been negligible. The factors making up a 'good' school are frequently rather nebulous or blindingly obvious and tautological (Ouston, 1998), consisting of items like an academic emphasis or high-quality leadership or good discipline. Reviews of the SEM exceed empirical studies, and the results of all studies are so far correlational rather than experimental in nature. The dangers of taking correlates as though they are the factors underlying school effects was recently illustrated in a spoof article presenting evidence that schools should be built on higher ground (Gorard *et al.*, 1998). SEM researchers are generally aware of this danger, but some have been found to move, almost unwittingly, from using factors as *descriptors* of successful schools, to referring to them as potential *determinants* of school improvement (in Hamilton, 1997). It is also the case that the long lists of correlated factors usually produced by SEM research sometimes contain apparently contradictory items. For example, one such list contains both 'strong firm leadership' and 'reciprocity'. Another list moves from effective leadership as exemplified by collegiality to the rigorous selection of teachers.

One unintended impact of SEM studies has been the apparent marginalisation of the role of important context factors. 'Adjusting for social factors has led some to a delusion that social factors don't matter' (Ouston, 1998, p. 7). The irony is that when school effectiveness models based on explaining the residual variance are used to push for small-scale policy changes like homework clubs or compulsory uniforms, their findings are often seized upon by ministers. When identical methods are used to point out the importance of value-added analyses, and the almost over-riding role of socio-economic context in school outcomes, then politicians are less happy. Tim Eggar (then an education minister) stated: 'we must not cover up underachievement with fiddled figures', while Michael Fallon (then Under-Secretary for Education and Science) claimed: 'we will not be dressing up the facts, obscuring the real level of performance by altering outcomes to take account of spurious measures of disadvantage or deprivation' (in Gipps, 1993). Despite these assertions, every study in this field has come to the conclusion that the role of context is paramount (e.g. Nuttall *et al.*, 1988). But there is no real sign that the government wishes to tackle the social inequality that lies at the heart of educational inequality, relying instead on school improvement (Hatcher, 1998).

> The problem of underperformance has been largely … conceived of as a failure of schools and of teachers … What School Effectiveness Research has failed to provide … is to develop an understanding of the processes which have led to the remarkably strong and surprisingly consistent relationship between socio-economic context and school performance. (Gibson & Asthana, 1998, p. 207).

Another, unintended, impact of these SEM studies may have been an exaggeration of the importance of examination results. While the researchers themselves have often been scrupulous in pointing out that results measure only some of the activities of a school, the fact that examinations appear easy to measure and monitor means that more complex school outcomes have been neglected. A good school has come to mean one with good exam results (Ouston, 1998), even though this is only a small part of what families look for when they choose a new school (Gorard, 1997).

> If schools are to be judged by examination results, there will be great pressure on schools to reflect this bias in their teaching. And the fact is that although … results are generally esteemed by parents and employers, they measure only a small part of what teachers would regard as desirable educational outcomes. They place a premium on propositional knowledge … so evidently does the public at large, to judge by the popularity of games of the 'mastermind' type. But what is more important is to use this knowledge procedurally … [and]… there still remains the whole area of a pupil's personal and social development which parents and employers rightly expect a school to foster (Holt, 1981, p. 18).

School effectiveness cannot be seen as a unitary trait applying to all subjects, departments, ages and abilities, and to both genders (Nuttall *et al.*, 1988), and schools may have both academic and non-academic effects. Schools that promote the first do not necessarily enhance personal and social development, for example (Smyth, 1998). While there is some indication that school effects may be consistent over time, they are not necessarily consistent over different outcomes or for different groups of students (Sammons *et al.*, 1996). They may have very limited and short-term impacts outside schools which do not, for example, carry over with the individual into university performance (Hughes *et al.*, 1996).

The Difficulties of International Comparisons

The next section begins to examine some of the more specific methodological difficulties in

carrying out international comparisons of educational systems and their outcomes, and concludes that such comparisons are fraught with problems.

Much of the so-called evidence which has been used to show that standards in Britain are falling, or that they are poor in international terms, can be dealt with very quickly. Much is hearsay, misinterpretation or 'academic Chinese whispers' (from Tooley & Darby, 1998). For example, when describing the early results of the SATs at KS1 and KS2, Phillips (1996) deplores the fact that many children failed to reach the standard in assessment that is expected for their age (see above). There have, of course, been disputes about the precise meaning of SAT results, but the reader can consider two possibilities to see how absurd Phillips' complaints are. If the 'standard for their age' is the minimum level that seven (or 11) year olds should reach, and in the first years of testing too many children did not reach that level, then who is to blame? The children or the target-setters? How is it possible to define a minimum standard for an age group without consideration of the *actual* standard of the age group? On what basis was the test calibrated? On the other hand, if the 'standard for their age' is an average figure, then a large number of children might be expected to achieve a lower grade, and an equal and opposite number might be expected to achieve a higher grade. The complaint is similar to those described by Huff (1991), where anxious parents worry that their child is not talking by the time of the average age for starting to talk. As Huff points out, around half of all parents should expect to be in this situation. It is what average means in this context.

This neglect of the quality of evidence (or in some cases lack of logic, where the comparator is missing) has been seen by some as endemic in the debate. Black (1992) for example says:

> Sweeping claims are made and repeated so often that the public come to accept them as self-evident truths. The outstanding example is the claim that standards have fallen. Such a claim cannot be supported by any review of the extensive evidence that such a sweeping generalisation must embrace. (p. 5)

An astounding example is recalled by Gipps (1993). In 1991, when the first KS1 SATs were held, the Secretary of State for Education announced to the press before the results were publicly available that nearly a third of all children aged seven were unable to recognise even three letters of the alphabet. If true, this would indeed have been newsworthy. It was not true, and the actual figure revealed a few days later was 2.5% of all children aged seven were unable to recognise three letters of the alphabet. The original story was, of course, the headline in all media reports on the day. No public retraction or correction was ever made. Presumably, many of the public still believe the original story to be true.

The findings from formal international studies of student performance are often more substantial, and worthy of greater attention. This is generally so for two main reasons. The researchers involved are academics (i.e. not politicians), and they generally face up to, rather than ignore, the difficult challenges of making systems comparable. The problems faced by researchers in this field include the comparability of different assessments, the comparability of the same assessments over time, using examinations or tests as indicators of performance at all, the different curricula in different countries, the different standards of record-keeping in different countries, and the competitiveness (especially) of developing countries (see O'Malley, 1998). Yet what international comparisons seek to do is solve not one but *all* and more of these problems at once. An observer who claims that on the basis of a standard test, one country has performed better than another, is also saying that the test involved similar children (it would not be fair to compare boys in one country with girls in another), who had followed a similar curriculum (it would not be fair to test people in a subject they had not

studied), that the test was a useful indicator of educational progress, and that it was administered in the same way in both countries. To use an extreme example to make the point, one would expect 16-year-old boys in Wales to have a better knowledge of the laws of the game of rugby than 11-year-old girls in Japan. Such a result would not make a useful international comparison.

Yet, in fact, most studies reported make precisely these types of 'unfair' comparison, although in less extreme (or perhaps more disguised) forms. One of the first such studies looked at maths results in Switzerland and compared them to those of Barking and Dagenham in Essex (in Brown, 1998). There are no prizes for guessing which 'country' came out on top. Students in Norway may look poor in assessments of their basic skills (see below), but then Scandinavia generally does not 'revel' in assessments as some countries do (OECD, 1996). The Norwegian EMIL project leads to a curriculum covering as broad a range of content as possible, and this breadth is what is assessed by them. As the OECD (1996) succinctly state: 'you get the school you test for' (p. 17). This may be partly why students in Singapore appear to do well in international comparisons of mathematics and science, since their assessment system favours progress in these areas in a 'lopsided' way, at least according to O'Malley (1998). When tests are not used, but the comparison is made between local qualification systems, it is almost impossible to decide on fair equivalencies for GCSE in Britain, the *baccalauréat* in France, and the *abitur* in Germany for example (Rafferty, 1999). There is therefore considerable potential for the 'fiddling' of figures by governments concerned to present a well-trained workforce to their potential overseas investors.

In Britain, at least, there is some balance in the use of fiddled figures. Although some sections of the government wish to argue that five GCSEs is equivalent to a *baccalauréat* and so elevate Britain's international ranking, there are other sections who ally with the media to present the more standard picture of education in crisis. Good news, such as that in the 1988 International Assessment of Educational Progress study in six countries, is therefore simply not reported. This study suggested that students in England and Wales excel in logic and problem-solving, beating even South Korea which excelled in almost everything else (Brown, 1998). A British review of several international studies by Reynolds & Farrell (1996) was used to press for the reintroduction of whole class teaching on a Taiwanese model, since results were consistently better in Taiwan. However, the BBC Panorama programme used to push this argument did not report that even in Taiwan there is the same 'long tail of underachievement' as in England. *The Independent* (11 June, 1997) reported that, the 'English came bottom of the class in Maths' in the TIMSS, although reading the piece in full revealed that England actually came tenth out of 17 countries mentioned. The same study reported that results from England were 'excellent' and improving in science and geometry (in Brown, 1998).

I have written elsewhere about the apparently poor performance of schools in Wales (e.g. Gorard, 2000c), which has been used to place the majority of comprehensive schools in Wales at the end of a very long chain of policy-borrowing (Gorard, 1998). British schools are supposed to model themselves on those from the Pacific Rim. Welsh schools are supposed to model themselves on those in England. English-medium comprehensives in Wales are supposed to model themselves on schools teaching through the medium of Welsh. One serious outcome of these comparisons is that they have a significant impact on the education that children experience. For example, the apparent evidence of the poor performance in mathematics of primary school students in Wales has led to a proposal for banning the use of calculators (Costley, 1999). Now such a change may be good, or it may be bad. The point here is that it *cannot* be justified by the research cited in evidence. Poor educational research

research can therefore lead to ill-thought changes in schools based on misunderstanding and misinformation (Gorard, 2000b).

Third International Mathematics and Science Study (TIMSS)

This section is devoted to consideration of perhaps the most convincing evidence for the relative failure of schools in England and Wales, the results of the TIMSS. The results for mathematics are shown in Table I. Sixteenth place for England is far from impressive, but better than several countries, including the United States, Norway and Spain. Many of the 16 other countries taking part, but not shown in Table I, also scored lower, but were omitted from analysis by the researchers as they did not meet the sampling requirements for the study. Of these 16, six did not meet the required participation rates, four did not meet the age limit requirements, three were judged to have poor sampling methods, and another three had more than one of these sampling problems. In this study of the attainment of 14 year olds, one South American country submitted scores for a cohort averaging 16 years of age.

Despite the necessary restrictions to samples imposed by the researchers, it is clear that Table I contains significant variation in the age of respondents. The oldest average age is for Singapore, at the top of the table in terms of score, and the youngest is for Iceland, near the bottom. The reasons for some of these differences are quite clear. Some countries allocate students to school years strictly by age (e.g. England), while other countries have yearly assessments and 'retake' years, leading to very different distributions of age per teaching class. In some comparisons the entire grade or year cohort is used in order to minimise disruption

TABLE I. Performance and mean age of top 23 countries in TIMSS

Country	Mathematics score	Mean age
Singapore	643	14.5
Korea	607	14.2
Japan	605	14.4
Hong Kong	588	14.2
Belgium	565	14.1
Czechoslovakia	564	14.4
Slovenia	547	14.3
Switzerland	545	14.2
France	538	14.3
Hungary	537	14.3
Russia	535	14.0
Ireland	527	14.4
Canada	527	14.1
Sweden	519	13.9
New Zealand	508	14.0
England	506	14.0
Norway	503	13.9
United States	500	14.2
Latvia	493	14.3
Spain	487	14.3
Iceland	487	13.6
Lithuania	477	14.3
Cyprus	474	13.7

Source: Beaton *et al.* (1996).

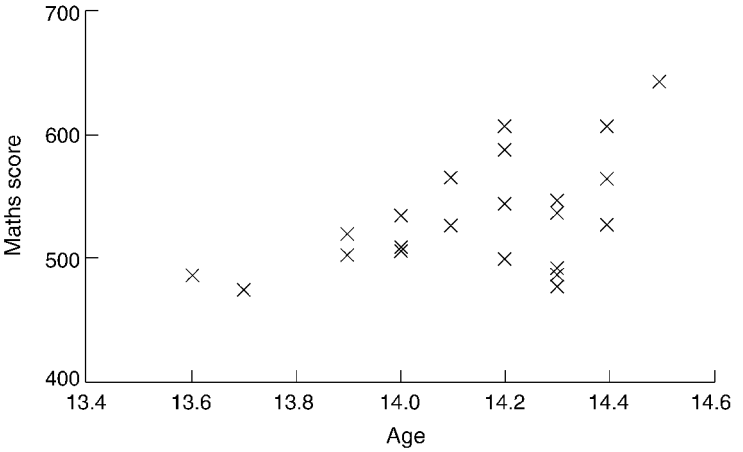

FIG. 1. Scatterplot of TIMSS maths score and mean age.

on the day of the test, so there have been reports of children aged 16 in Germany taking part in tests meant for 13 year olds (Brown, 1998). The linear correlation between age and score in Table I is + 0.53 (Pearson's r, significant at the 1% level). Figure 1 shows this relationship graphically. This means that one would expect countries with older children in the test to have higher scores, and that nearly 30% of the variance in outcomes is explicable by differences in mean age alone.

There may be further problems with the sampling, unchecked by the guidelines used by the researchers. Some countries have special schools for those with special educational needs, who would not therefore take part, while others have a policy of integration. Many countries, including Singapore, do not have compulsory schooling for all, so that in Thailand for example only around 32% of the relevant age group go to school. Attendance rates matter since only those students in school took part, and there is no reason to assume that those not attending school would have performed at the same level. In England some LEAs refused to participate in TIMSS from the start. Of the 300 schools that were asked to participate, 162 agreed, and for some parts of the study the response rate from these 162 schools was as low as 87% (Keys *et al.*, 1996a). The implication is that, for England, the overall response rate was below 50%. One can only imagine what the equivalent participation data would be for many other countries. It is partly for these reasons that Brown (1998) concludes that the information in international league tables is generally too flawed to be of any use at all.

Even if the samples used were good random samples of each country, which they clearly were not, the results would still have a standard sampling error. On this very generous assumption, the sampling error would mean that any population statistic, such as a mean score, is 95% likely to lie within two standard deviations of the corresponding sample statistic (Keys *et al.*, 1996a, p. 47). Using this as a guide shows that only the Russian Federation and Sweden (of the countries meeting the sample criteria) had an equivalent or younger mean age than students in England, *and* a higher mathematics score without 95% confidence intervals overlapping with those of England. Singapore, for example, has an average age of 14.5 versus 14.0 for England. New Zealand had the same mean age as England, but although its score is two points (0.4%) higher than England, this score has a standard deviation of more than twice that difference (Keys *et al.*, 1996b, Table I). Therefore the probability is quite high that at least some of the scores appearing to be greater than those from England are taken from a population whose mean level is actually lower than England's (and vice versa of course).

One reason that mathematics and science are used in comparisons is that they are felt to be more culture-free than other school subjects (such as English language). Even so, on close examination the variation in what had been taught to children in each country by the age of 14 was considerable. The curriculum in England had only covered an estimated 57% of the content of the TIMSS test by age 14, leading to the situation of children being tested on material they had not learnt. They had, instead, covered other topics in mathematics which were not in the test. Children from the the United States had the best match between curriculum and test material, perhaps because the the United States were the primary funders of the study and therefore created the test. It is interesting to note that children in England still scored higher than those in the United States.

At least part of the reason for this apparently poor showing by the United States may be due to motivational factors. American students were reportedly held back from a games lesson to take part, and told that the test did not count towards their grades. In South Korea, on the other hand, the children were urged to do well for the sake of their country, and marched in to the sound of school bands. If generally true, these systematic differences are sufficient in themselves to negate the generally small differences between many scores, which are obscured by the use of ranks in reporting the outcomes. In the same way, the use of mean scores alone hides the variation within each country. This within-country variation is usually similar, except that many of the more 'successful' Pacific Rim countries have a higher standard deviation than England, signifying more varied outcomes, and a longer 'tail' below the mean.

Other International Indicators

Of the 23 countries in an OECD comparison in 1992, only Germany, Norway, Switzerland and the United States had a clearly higher proportion of their population than the UK educated to upper secondary level (Table II), and the situation remained the same in 1996 and is predicted to remain so until at least 2015 (CERI, 1998). It was only at tertiary level that the UK fell markedly behind many OECD 'competitors' (OECD, 1993). By 1996, this deficit had been corrected according to the figures, and the net entry rate for university-level education was 41% in the UK, the fourth highest of 18 countries in the study (CERI, 1998). In 1996, the UK had one of the largest number of 'expected' years of education, and the third highest ratio of university-level graduates to population, along with perhaps the most balanced figures for participation by gender at all levels of initial education.

Although there are systematic inequalities in participation and qualifications in the UK, these are generally smaller than in other OECD countries and diminishing over time. For example, of the 15 EU countries only three have a smaller proportion of 16–19 year olds in full-time education from families with parents educated only to upper-secondary level. The UK has the lowest proportion of 16–19 year olds in full-time education from families with parents educated only to lower-secondary level (Eurostat, 1998). While all of these figures need to be treated with the same caution applied to other sources in this article, and indeed OECD stress the difficulties and compromises involved in making international comparisons of this sort, there is no evidence from successive OECD and EU figures since 1985 that initial education in the UK is either poor or inequitable in international terms.

Was There a Golden Age?

Progress in the 20th century has led to considerable improvements in social inclusion and opportunities by gender, ethnicity and class, and these improvements apply to education as

TABLE II. Highest education level of those aged 25–64

1992	Upper secondary	Tertiary (university and HE)
United States	84.1	30.2
Germany	81.9	21.4
Switzerland	80.7	20.9
Norway	78.9	25.2
Canada	71.3	41.1
Sweden	69.9	24.1
UK	68.0	18.4
Austria	67.9	6.9
Finland	62.1	19.2
Denmark	58.9	19.2
Netherlands	57.8	20.9
New Zealand	56.4	23.6
Australia	52.8	22.4
France	52.2	15.9
Belgium	45.2	20.2
Ireland	42.2	16.9
Italy	28.5	6.4
Spain	22.9	13.0
Portugal	14.2	6.7
Turkey	13.7	4.8

Source: CERI, 1998.

much as any other social phenomenon. It would be inappropriate to deny, or downplay, this progress. Whatever our complaints may be in retrospect, the 1944 Education Act, which introduced secondary education for all, the comprehensivisation of schools, perhaps even the 1988 Education Reform Act, as well as a host of other initiatives, have all attempted to produce greater social justice in the English education system, and to some extent they have succeeded. To admit these improvements is not to deny the existence of the remaining problems, but to help describe the current situation more precisely and so define those problems more closely.

Reasonably authoritative sources can be cited to show that people of all periods in the 20th century were worried about standards, underachievement, literacy and inequality in education pretty much as they are now. This is certainly true of 1904 and 1943 (in Lestor, 1979). There may be a tendency for each generation to perceive a decline in standards. Perhaps it is simply that more and more is expected from education over time.

In fact, some evidence is available that schools are not only improving their statutory assessment scores over time, but that other problems, such as truancy, are declining, while schools are becoming more efficient in economic terms (Bradley *et al.*, 1999). If true, it is welcome that these general improvements in the standards of, and outcomes from, education also appear to be reducing the educational inequalities between different social groups and geographical regions. Kelsall & Kelsall (1974) present some evidence that the gap between the top and bottom of the social scale in economic, power and status terms was being reduced by the 1970s. Although inequality and injustice for the socially disadvantaged has always existed (MacKay, 1999), in fact, 'if you take a long-term historical perspective of the provision of education in the UK throughout its entire statutory period … you could say that a constant move towards greater justice and equity has been the hallmark of the whole process' (p. 344).

Conclusion

The thrust of this article has been to suggest that a consideration of the standards or effectiveness of a school system is not a simple matter of counting and comparison. Even where simplifying assumptions are made about the outcomes from schools, such as a concentration on statutory assessment and test results, philosophical and methodological difficulties persist. In fact, it is sometimes difficult to discover what difference schools actually make to attainment even in these very restrictive terms (and, of course, success can refer to many things other than examination scores). In the light of these difficulties, there is certainly no evidence here of falling educational standards over time in Britain, and no convincing evidence of underperformance relative to the educational systems of other developed nations. Perhaps neither component of the crisis account has much basis in fact.

If the argument thus far is accepted there remains very little solid evidence to create an evidence-base for international policy borrowing. Issues of the reliability and validity of assessments over time and place within one country become still more complex when test scores from more than one country are to be compared. The growing use of international comparisons, a fascinating area in its own right, does therefore lead to two potential problems. First, the difficulty of comparing complex systems seems to encourage a concentration on the 'lowest common denominator' of assessment, and therefore furthers an undue focus on credentials. Second, the lack of consistency in what is being measured even in these limited terms means that naïve observers may be tempted to draw unwarranted conclusions about the relative merits of entire national school systems.

REFERENCES

Barber, M. (1996) *The Learning Game* (London, Indigo).

Beaton, A, Mullis, I., Martin, M., Gonzalez, E., Kelly, D. & Smith, T. (1996) *Mathematics Achievement in the Middle School Years: IEA's Third Mathematics and Science Study* (Chestnut Hill, Massachusetts, Boston College).

Bentley, T. (1998) *Learning Beyond the Classroom* (London, Routledge).

Black, P. (1992) *Education: putting the record straight* (Stafford, Network Educational Press).

Bolton, E. (1993) Imaginary gardens with real toads, in: C. Chitty, & B. Simon (Eds) *Education Answers Back*, pp. 3–16 (London, Lawrence & Wishart).

Boyson, R. (1975) *The Crisis in Education* (London, Woburn Press).

Bradley, S., Johnes, G. & Millington, J. (1999) *School Choice, Competition and the Efficiency of Secondary Schools in England* (Management School, Lancaster University (mimeo)).

Brown, M. (1998) The tyranny of the international horse race, in: R. Slee, G. Weiner & S. Tomlinson (Eds) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements,* pp. 33–47 (London, Falmer Press).

Cassidy, S. (1999a) Pass mark 'fiddle' is strenuously denied, *Times Educational Supplement*, 28 May, p. 2.

Cassidy, S. (1999b) English papers at 14 were 'too hard', *Times Educational Supplement*, 4 June, p. 6.

Cassidy, S. (1999c) Test scores did not add up, *Times Educational Supplement*, 16 July, p. 5.

CERI (1998) *Education at a Glance: OECD indicators* (Paris, OECD).

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfield, F. & York, R. (1966) *Equality of Educational Opportunity* (Washington, DC, US Government Printing Office).

Corbett, A. (1993) The peculiarity of the English, in: C. Chitty & B. Simon, (Eds) *Education Answers Back,* pp. 17–30 (London, Lawrence & Wishart).

Costley, M. (1999) Hain calls for school curb on calculators, *Western Mail*, 11 January, p. 1.

Cox, C. & Dyson, A. (1990) An open letter to Members of Parliament, in: B. Moon, J. Isaac & J. Powney (Eds) *Judging Standards and Effectiveness in Education*, pp. 103–109 (London, Hodder & Stoughton).

Creemers, B. (1994) The history, value and purpose of school effectiveness studies, in: D. Reynolds, B. Creemers, P. Nesselradt, E. Shaffer, S. Stringfield & C. Teddlie (Eds) *Advances in School Effectiveness Research and Practice*, pp. 2–17 (Oxford, Pergamon).

Cresswell, M. & Gubb, J. (1990) The Second International Maths Study in England and Wales: comparisons between 1981 and 1964, in: B. Moon, J. Isaac & J. Powney (Eds) *Judging Standards and Effectiveness in Education* (London, Hodder & Stoughton).

DALY, P. (1991) How large are secondary school effects in Northern Ireland?, *School Effectiveness and School Improvement*, 2 (4), pp. 305–323.

DAUGHERTY, R. (1995) *National Curriculum Assessment: a review of policy, 1987–1994* (London, Falmer Press).

DfEE (1999) *A Fresh Start: improving literacy and numeracy* (Nottingham, DfEE Publications).

ETAG (1998) *An Education and Training Plan for Wales* (Cardiff, Education and Training Action Group).

EUROSTAT (1998) *Social Portrait of Europe September 1998* (Brussels, Statistical Office of the European Communities).

GIBSON, A. & ASTHANA, S. (1998) School performance, school effectiveness and the 1997 White Paper, *Oxford Review of Education*, 24 (2), pp. 195–210.

GIPPS, C. (1993) Policy-making and the use and misuse of evidence, in: C. CHITTY & B. SIMON (Eds) *Education Answers Back*, pp. 31–44 (London, Lawrence and Wishart).

GORARD, S. (1997) *School Choice in an Established Market* (Aldershot, Ashgate).

GORARD, S. (1998) In defence of local comprehensive schools in South Wales, *Forum*, 40 (2), pp. 58–59.

GORARD, S. (2000a) Questioning the crisis account: a review of evidence for increasing polarisation in schools, *Educational Research*, 42 (3), pp. 309–321.

GORARD, S. (2000b) *Education and Social Justice* (Cardiff, University of Wales Press).

GORARD, S. (2000c) 'Underachievement' is still an ugly word: reconsidering the relative effectiveness of schools in England and Wales, *Journal of Education Policy*, 15 (5), pp. 559–573.

GORARD, S., REES, G. & JEPHCOTE, M. (1998) The role of contour lines in school improvement, *Research Intelligence*, 66, pp. 30–31.

GRAY, J. & WILCOX, B. (1995) *'Good school, Bad School': evaluating performance and encouraging improvement* (Buckingham, Open University Press).

HAMILTON, D. (1997) Peddling feel-good fictions, in: J. WHITE & M. BARBER (Eds) *Perspectives on School Effectiveness and School Improvement*, pp. 27–43 (London, Institute of Education).

HAMILTON, D. (1998) The idols of the market place, in: R. SLEE, G. WEINER & S. TOMLINSON (Eds) *School Effectiveness for Whom? Challenges to the school effectiveness and school improvement movements*, pp. 13–20 (London, Falmer Press).

HARKER, R. & NASH, R. (1996) Academic outcomes and school effectiveness: Type 'A' and Type 'B' effects, *New Zealand Journal of Educational Studies*, 32 (2), pp. 143–170.

HATCHER, R. (1998) Labour, official school improvement and equality, *Journal of Education Policy*, 13 (4), pp. 485–499.

HOLT, M. (1981) *Evaluating the Evaluators* (London, Hodder & Stoughton).

HUFF, D. (1991) *How to Lie with Statistics* (Harmondsworth, Penguin).

HUGHES, D., LAUDER, H. & STRATHDEE, R. (1996) The short term limits to school effectiveness studies, a reply to Harker, *New Zealand Journal of Educational Studies*, 31 (2), pp. 199–201.

IEA (1988) *Science Achievement in Seventeen Countries: a preliminary report* (Oxford, Pergamon).

JENCKS, C., SMITH, M., ACKLAND, H., BANE, M., COHEN, D., GINTIS, H., HEYNS, B. & NICHOLSON, S. (1972) *Inequality: assessment of the effect of family and schooling in America* (New York, Basic Books).

KELSALL, R. & KELSALL, H. (1974) *Stratification* (London, Longman).

KEYS, W., HARRIS, S. & FERNANDES, C. (1996a) *Third International Mathematics and Science Study: National Report appendices* (Slough, NFER).

KEYS, W., HARRIS, S. & FERNANDES, C. (1996b) *Third International Mathematics and Science Study: First National Report Part 1* (Slough, NFER).

KITCHEN, A. (1999) The changing profile of entrants to mathematics at A-level and to mathematical subjects in higher education, *British Educational Research Journal*, 25 (1), pp. 57–74.

LAUDER, H., HUGHES, D., WATSON, S., WASLANDER, S., THRUPP, M., STRATHDEE, R., SIMIYU, I., DUPUIS, A., MCGLINN, J. & HAMLIN, J. (1999) *Trading in Futures: why markets in education don't work* (Buckingham, Open University Press).

LESTOR, J. (1979) Was there a golden age?, in: H. PLUCKROSE & P. WILBY (Eds) *The Condition of English Schooling*, pp. 123–139. (Harmondsworth, Penguin).

MACKAY, T. (1999) Education and the disadvantaged: is there any justice?, *The Psychologist*, 12 (7), pp. 344–349.

NATIONAL COMMISSION ON EDUCATION (1993) *Learning to Succeed* (London, Heinemann).

NOAH, H. & ECKSTEIN, M. (1992) Comparing secondary school leaving examinations, in: M. ECKSTEIN & H. NOAH (Eds) *Examinations: comparative and international studies*, pp. 3–24 (Oxford, Pergamon Press).

NUTTALL, D. (1979) The myth of comparability, *Journal of the National Association of Inspectors and Advisers*, 11, pp. 16–18.

NUTTALL, D. (1987) The validity of assessments, *European Journal of the Psychology of Education*, II (2), pp. 109–118.

NUTTALL, D., GOLDSTEIN, H., PRESSER, R. & RASBASH, H. (1988) Differential school effectiveness, *International Journal of Educational Research*, 13 (7), pp. 769–776.

OECD (1993) *OECD Education Statistics, 1985–1992* (Paris, OECD).
OECD (1995) *Schools Under Scrutiny* (Paris, OECD).
OECD (1996) *Evaluating and Reforming Education Systems* (Paris, OECD).
O'Malley, B. (1998) Measuring a moving target, *Times Educational Supplement*, 18 September, p. 22.
Ouston, J. (1998) The school effectiveness and improvement movement: a reflection on its contribution to the development of good schools. Paper presented at ESRC Redefining Education Management seminar, Open University, 4 June.
Petre, J. (1999) Education standards in decline, says survey, *Sunday Telegraph*, 25 July, p. 13.
Phillips, M. (1996) *All Must Have Prizes* (London, Little, Brown & Co.).
Porter, M. (1990) *The Competitive Advantage of Nations* (London, Macmillan).
Postlethwaite, T. (1985) The bottom half in secondary schooling, in: G. Worswick (Ed.) *Education and Economic Performance*, pp. 93–100 (London:, NIESR).
Prais, S. (1990) Mathematical attainments: comparisons of Japanese and English schooling, in: B. Moon, J. Isaac & J. Powney (Eds) *Judging Standards and Effectiveness in Education*, pp. 65–83 (London, Hodder & Stoughton).
Rafferty, F. (1999) UK teenagers through exam mill, *Times Educational Supplement*, 30 July, p. 1.
Reynolds, D. (1990) School effectiveness and school improvement: a review of the British literature, in: B. Moon, J. Isaac, & J. Powney (Eds) *Judging Standards and Effectiveness in Education*, pp. 17–37 (London, Hodder & Stoughton).
Reynolds, D. & Farrell, S. (1996) *Worlds Apart? A review of international surveys of educational achievement involving England* (London, HMSO).
Sammons, P. Mortimore, P. & Thomas, S. (1996) Do schools perform consistently across outcomes and areas?, in: J. Gray, D. Reynolds, C. Fitz-Gibbon & D. Jesson (Eds) *Merging Traditions: the future of research on school effectiveness and school improvement*, pp. 3–26 (London, Cassell).
Schagen, I. & Morrison, J. (1998) *QUASE Quantitative Analysis for Self-evaluation. Overview report 1997: Analysis of GCSE cohorts 1994 to 1996* (Slough, NFER).
Shipman, M. (1997) *The Limitations of Social Research* (4th edition) (Harlow, Longman).
Skills and Enterprise Briefing (1999) *Skills and Enterprise Briefing Issue 5/99* (Sudbury, DfEE Publications).
Skills and Enterprise Network (1999) *Skills and Enterprise Network Annual Conference Report* (Sudbury, DfEE Publications).
Smyth, E. (1998) School effectiveness in the Republic of Ireland: a multidimensional analysis. Presentation to BERA Annual Conference, Belfast.
Stoll, L. & Fink, L. (1996) *Changing Our Schools: linking school effectiveness and school improvement* (Buckingham, Open University Press).
Tooley, J. & Darby, D. (1998) *Educational Research: a critique* (London, OFSTED).
Woodhead, C. (1999) Education was a lottery … , *News of the World*, 25 July, p. 16.