# Assessment, League Tables and School Effectiveness: Consider the Issues and 'Let's Get Real'!

**Kenneth J. Rowe**
*The Australian Council for Educational Research, Australia*

## Abstract

Current policy activity related to 'outcomes-based' educational performance indicators and its link with growing demands for accountability, standards monitoring, target-setting, benchmarking and school effectiveness is widespread – in Australia and internationally. Within this context, the present paper highlights the limitations of using performance indicators based on test or examination scores as accountability measures at the school- and system-level, or indeed, as measures of student learning outcomes. The issues raised are presented for consideration, stressing the need for caution in generating and publishing potentially invalid and misleading information, especially in the typically published form of 'league tables' consisting of schools' raw, 'ability-adjusted', or 'value-added' average achievement scores, with the risk of generating both individual and institutional harm. As a means of at least minimising such problems, the paper outlines a code of ethics for the publication of educational performance indicators along the lines proposed by Goldstein and Myers (1996), and Myers and Goldstein (1996).

## The context

In a discussion of the relationship between assessment and learning, Broadfoot (1996) claims: 'Assessment is arguably the most powerful policy tool in education' (p. 21). Anyone observing the protracted political and media furore in Britain during January-February 1996, following the release of: (a) the 1994-1995 national curriculum test results in English, mathematics and science for 7, 11 and 14 year-olds,[i] and (b) the ensuing Office for Standards in Education report with its emphasis

Submitted for publication June 26, 2000; final revision received August 21, 2000; accepted August 25, 2000.

Contact: rowek@acer.edu.au

Citation: Rowe, K. (2000) Assessment, League Tables and School Effectiveness: Consider the Issues and 'Let's Get Real'! *Journal of Educational Enquiry,* Vol. 1, No. 1, pp. 73-98.
[URL: http://www.education.unisa.edu.au/JEE]

on 'teacher incompetence',[ii] could be forgiven for accusing Broadfoot of gross understatement! Rather, '...current educational policy activity related to issues of accountability, assessment, standards monitoring, performance indicators, quality assurance and school effectiveness is widespread' (Rowe, Hill & Holmes-Smith, 1995, p. 217). Moreover, they occupy front and centre of the political and media stage with persistent regularity. Since schooling accounts for a significant proportion of public and private expenditure, as well as generating a substantial quantity of paid employment for teachers and administrators, the enduring interest by governments and the media in the relative effectiveness of school education and its improvement is not surprising (see Fitz-Gibbon, 1996; Hill, 1995; Mortimore, 1998; Tucker & Codding, 1998; Tuijnman & Postlethwaite, 1994; Rowe, 2000a; Watson, 1996).

This is an especially sensitive issue at the present time given the level of consensus regarding the importance of school education as an element of micro-economic reform and in meeting the constantly changing demands of the modern workplace (NCEE, 1997). Despite the difficulties entailed in defining an *effective school* (Mortimore, 1991, 1998) and reaching consensus on the relevant criteria (Scheerens & Bosker, 1997), a good deal of discussion has focused on what is meant by *quality schooling* (eg., Chapman, Angus, Burke & Wilkinson, 1991) and how it might be measured (eg., Cuttance, 1990; Fitz-Gibbon, 1996; Masters, 1990, 1991, 1994; Rowe & Hill, 1998; Watson, 1996). Although the term *quality* is likewise problematic (see Istance and Lowe, 1991), the '...measurement of the *quality* of schooling is of critical importance at a time when so much school reform in so many parts of the world is being undertaken' (Mortimore, 1991, p. 214). In fact, concerns about the *quality* of school education and its monitoring have long been high priority policy issues in all OECD countries (OECD, 1989; 1993).

During the last decade, education systems throughout the world have been exposed to considerable reform and change – all justified on the grounds (or at least the rhetoric) of improving the *quality* of school education (see Harlen, 1994; Marsh, 1999; Mortimore, 1998; Tuijnman & Postlethwaite, 1994). A key feature of this change has been the frequent revisions of style and policy, especially in the area of accountability involving the assessment and monitoring of student learning outcomes. In the United States, detailed public accountability of schools and State education systems on the basis of students' test scores is well established, despite vigorous debate about the consequences of basing performance indicator and accountability arrangements solely on the outcomes of system-wide, standardized testing/ assessment programs. The debate has ranged from concerns about the negative impact of such programs on curriculum, teaching and learning (eg., Smith, 1991; Smith & Rottenburg, 1991), to the way in which results at State level have been manipulated for essentially political ends (eg., Cannell, 1988).

In Australia, policy revisions have been evident in the increasing national approaches to educational governance and accountability (Marginson, 1993), first signalled in the paper entitled *Strengthening Australia's Schools* (Dawkins, 1988) which called for a national focus on student assessment and standards monitoring. In commenting on this change, Wilson (1996) observed: 'One of the remarkable

things that has happened in Australian education in recent years is the rapid achievement of a hegemony by the idea of *outcomes-based education'* (p. 5). Wilson supports this observation by pointing to two features of this hegemony: (1) the considerable degree of national consensus that has been achieved about the purposes of education, and (2) that the consensus is based on student learning outcomes. The emergence of specific learning outcome statements in documents such as the Victorian *Curriculum and Standards Framework* (Board of Studies, 1994, 1999), and similar variants in other States and Territories, reflect the national influence in the work done on profiles (Hill, 1994; Rowe & Hill, 1996; Wilson, 1993) and strong commitments to *standards* and *outcomes based education* (Donnelly, 1999; Hannan, 1995; Hill & Crévola, 1999). These commitments are likewise reflected in recent increases in State-wide assessment and monitoring programs.[iii] In this context, Hill (1995, p. 4) notes:

> ...accountability pressures have forced most education systems to press ahead with large-scale assessment programs. All government school education systems in Australia, except the ACT, now operate programs to monitor educational standards. ... The principal motivation behind current assessment programs is to meet public demands for educational systems to be accountable for maintaining and indeed improving standards. As such, they tend to command broad support from the community, but rarely receive enthusiastic support from the teaching profession.

In the United Kingdom, recent government policies centred on *educational accountability* and *standards monitoring*, have had notable impacts on schools. Foremost among these is the implementation of a national curriculum, national testing, a new external school inspection system administered by the *Office for Standards in Education*, and the publication of schools' average achievement scores on tests and public examinations. As part of a general policy initiative by the British government since 1987 to promote the use of indicators by which public service institutions can be compared and their performances evaluated, the *Parents' Charter* (DES, 1991), for example, requires that comparative 'league tables' of examination and national curriculum test results be published for every educational institution (schools) and local education authority (LEA's). The 'league tables' consist of schools' rankings computed from students' average achievement scores (raw and unadjusted) on national curriculum test results at ages 7, 11 and 14 years, together with similar scores for the *General Certificate of School Education* (16 year-olds) and *A-levels* (18 year-olds). The stated intention of the *Parents' Charter* is that these tables be used by parents to assist in choosing schools for attendance by their children. However, consistent with Hill's point quoted above, the British government's intention in pursuing these policies is to meet public demands for accountability and the maintenance of educational standards.

The regnant 'market ideologies' that underpin such policies have fostered a climate in which competition has begun to dominate co-operation. The focus on allowing market forces to predominate makes it possible for governments and educational regulatory bodies to locate blame for 'poor performance' or 'ineffectiveness' at the local and/or school level. Since markets operate through competition in which there are 'winners' and 'losers', the designation of schools as 'effective' or 'ineffective' is seen as an inevitable consequence. The results for the

recipients of a 'failing'/'ineffective' label can be catastrophic. They may simply go out of business; they may be taken under the direct control of a state education department or, as happened recently in the UK, a 'failing' school was investigated by a government appointed commission and subsequently closed. From recent UK experience (see Goldstein, 1997a,b,c; Goldstein & Cuttance, 1988; Myers & Goldstein, 1996), the impact of 'league tables' has been evident in:

- Political and media 'bashing' of schools and teachers (see Notes i and ii).
- A test-dominated curriculum (particularly in English, mathematics and science) that has resulted in an over-emphasis (exclusive in some cases) on curriculum content that is to be tested or examined.
- Overt lobbying of the government by principals of non-selective schools to 'select' up to twenty per cent of their school enrolments in an attempt to improve their schools' rankings on the 'league tables'. This has resulted in a reluctance, and in some cases, direct refusals to enrol 'low-achievers'. Further, some schools have responded by concentrating their efforts on those students considered capable of improving their average examination and test scores, while giving less attention to those perceived less likely to improve.
- Parents have 'voted with their feet' by choosing to enrol their children in schools on the basis of 'league table' rankings. In some cases, this has meant changing their former residential locations to those in closer proximity to the chosen schools.

In any event, an inevitable result of comparisons among schools, whether by publication of crude 'league tables' as in the UK and in several Australian states,[iv] or more sophisticated 'value-added' ones like those published in the US State of Tennessee (Sanders & Horn, 1994), is that there are 'winners' and 'losers' (see Saunders, 1999). Once the losers are deemed to be 'failing' or 'ineffective' it is difficult to find ways of helping them in a prevailing social and political atmosphere of blame, recrimination and retribution. Whereas there are long-standing socio-political and socio-economic dissimilarities between education systems in Australia, Europe, the UK and US – particularly in respect of parental choice of schooling for their children – current policy emphases in Australia on educational accountability and the public dissemination of school performance indicator information have moved closer to international trends. Indeed, '…Australian politicians and senior bureaucrats currently advocating the publication of such performance information in the form of 'league tables', are naively, and in typical fashion, stomping around in an uninformed epistemopathological fog' (Rowe, 2000b, p. 46).

Given this context, the purpose of the present paper is to highlight the limitations of using educational performance indicators in the form of test or examination scores as instruments of accountability at the school and system level, or as 'measures' of student learning outcomes. This is not to deny the utility of such measures for diagnostic and student/school improvement purposes, or indeed, the obvious benefits arising from open access to accurate and reliable information (see Rowe, 1999b; Rowe, Turner & Lane, 2000). Rather, the issues raised here are presented for consideration, stressing the need for caution in generating potentially

invalid and misleading information with the potential risk of individual and institutional harm. The issues raised draw heavily on the discussions presented in several published papers, namely, Goldstein (1997a,b,c), Goldstein and Myers (1996), Goldstein and Spiegelhalter (1996), Goldstein and Thomas (1996), Myers and Goldstein (1996), Rowe (1996a,b, 1999a,b,c, 2000a,b), Rowe and Hill (1996, 1998), Rowe, Turner and Lane (2000), and Watson (1996).

## The assessment-accountability dilemma

Whereas the long-term goals of school education may be expressed as the enhancement of young peoples' access to and participation in society, as well as preparation for meeting the constantly changing demands of the modern workplace (OECD, 1983, 1986), the most direct and readily accessible measures of schooling outcomes are obtained from assessments of students' academic attainments. Herein, however, lies a dilemma that is evidenced in strident critiques of traditional and prevailing psychometric models for test and examination modes of assessment (eg., Berlak, 1992) and an equally strident chorus of concern for the deleterious effects of test-driven and test-dominated curricula (eg., Kellaghan, Madaus & Airasian, 1992). As Watson (1996) notes: 'In high stakes testing environments, educational practitioners are likely to distort their behaviour in order to meet the demands of the indicator, usually to the detriment of their real job' (p. 119). Nisbet (1993, p. 25) further highlights this dilemma in the following terms:

> In today's schools, assessment is a main influence on how pupils learn and how teachers teach. Whether assessment is in the form of examinations and tests, or marks and grades for coursework, its influence is pervasive. Often it distorts the process of learning through teaching to the test, cramming, short-term memorising, anxiety and stress – to the extent that learning to cope with assessment has become almost as important as the genuine learning which such assessments are supposed to measure. For many young people, assessment dominates education.

To date, for the purposes of determining educational and occupational access, standards monitoring, performance indicators, benchmarking, target-setting, accountability, and school effectiveness, the *measurement* of learning outcomes at the student, school, system, national and international levels has relied almost exclusively on the use of standardised achievement tests or public examinations (Goldstein & Lewis, 1996; Scheerens, 1993; 1995; Thomas et al., 1997). Robert Wood's (1986) comment that, 'In Britain, to talk of educational measurement is to talk of examinations' (p. 197), continues to hold as it has for the past fifty years. In the United States especially, the use of standardised achievement tests, dominated by the psychometric technology of item response modelling, has prevailed over the same period. Although the use of such tests and examinations for the measurement and evaluation of *educational effectiveness* is typically justified on the grounds of maximising *reliability* and ensuring *comparability*, it is argued cogently that this has been mostly at the expense of *validity* (Broadfoot, 1996; Lacey & Lawton, 1981; Moss, 1994). For example, in summarising the British, European and North American attempts at curriculum and assessment reform during the 1970's, Lacey and Lawton (1981, pp. 229-230) warned:

> ...conventional standardised achievement tests have inherent risks as instruments of evaluation for accountability since they seldom cover more than the common core or very basic curriculum units. Thus, as the sole instrument, they may be highly deceptive because of lacking content validity. ...test scores as such have low information value about the outlying processes as well as the environmental and administrative frame conditions necessary to understand and appreciate the skills and efforts needed to fulfil a certain educational goal.

Problems of content validity, however, would appear to be less acute in studies that have made use of public examination results, such as the study reported by Tymms (1993), since public examinations are designed to assess learning outcomes as set out in some detail in syllabi on the basis of which it can be assumed that teachers and schools have followed closely. Where examination scores have been used as outcome measures, differences between classes and faculties within schools are typically large and substantially greater than differences among schools, although effects are not especially consistent across faculties or from year to year.

From the U.S. there has long been criticism of the utility of standardised tests as measures of either learning or competence (see Darling-Hammond, 1994; Newmann & Archbald, 1990). Newmann and Archbald argue, for example, that '...most data currently used to assess schools' performance, especially scores on standardised tests, fail to measure meaningful forms of human competence and that significantly new forms of assessment need to be developed' (p.164). From Britain, the use of traditional tests and examinations for monitoring and accountability purposes has likewise been the focus of intense critical discussion and calls for alternatives (see Broadfoot, 1996; Gipps & Murphy, 1994; Murphy & Broadfoot, 1995).

For a variety of epistemological and methodological reasons such criticism has been manifestly ignored in almost all large-scale monitoring exercises of student learning outcomes. This is most notable in the growing field of school effectiveness research (see Hill & Rowe, 1996. 1998; Mortimore, 1998; Reezigt, Guldemond & Creemers, 1999; Rowe & Hill, 1998; Rowe et al., 1995; Rowe & Rowe, 1999; Scheerens & Bosker, 1997; Thomas et al., 1997), where the identification of 'effective schools' continues to be made on the basis of limited operational definitions of what it means to be a 'good school'. Applying an apparent embargo on measures of presumably desirable affective and behavioural outcomes of schooling, for example, the most common approach is to identify those schools with aggregated students' scores on standardised tests of reading and mathematics (or on examinations) that are higher than average, regardless of the curriculum validity of such measures. In so doing, there seems to be little awareness that '...the majority of such tests assess skills in terms of generalised academic *abilities* and enduring cognitive 'traits' rather than specific learning outcomes arising from classroom instruction' (Hill & Rowe, 1996, p. 7). Under such circumstances, claims for 'school effectiveness' *per se*, are at best, tenuous.

In several school effectiveness studies, the problem of the content validity of items contained in standardised tests has been recognised and attempts have been made to assess the extent to which students have had the opportunity to learn the content represented in individual test items. Where this has been done, it has

frequently been observed that 'opportunity to learn' is a major explanation for patterns of performance on the tests. This is true in the case of the study reported by Bosker, Kremers and Lugthart (1990), and has been a consistent finding in various international studies of achievement (see Bosker & Scheerens, 1989, 1994; Scheerens & Bosker, 1997; Hill & Rowe, 1998). Unfortunately, it is not always possible to determine whether lack of 'opportunity to learn' reflects unsatisfactory test validity or inadequate instruction.

Nevertheless, elements of this criticism have gained credence in the areas of standards monitoring and performance assessment in the U.S. where alternative approaches to obtaining more curriculum-specific and 'authentic' (Wiggins, 1989) measures have been considered (eg., Floden, 1994; Shavelson, 1994; Taylor, 1994). Similarly, the work of Broadfoot (1994), Gipps (1996) and Gipps and Murphy (1994) is representative of a growing number of U.K. educationalists who have called into question prevailing 'mechanistic', 'objectives-driven' modes of assessment, and re-assert the teacher's professional role in education by challenging the widespread assumption that teachers' assessments are less reliable than those obtained from examinations and tests. For a discussion of related developments in recent years throughout the world, including those in Australia, see Rowe and Hill (1996).

## Limitations with the use of test and/or examination scores as performance indicators for evaluating school effectiveness

In proposing the development of a 'performance indicators framework for schooling' in Australia, Watson (1996, p. 110) notes: 'It is important to acknowledge the limitations of a performance indicators framework for evaluating school effectiveness'. While this is true as a general proposition (see discussion below), it is not sufficient to merely invoke the inherent '..complexity of service provision in social policy areas such as education...' (ibid.) due to the fact that '...school education...has multiple objectives, multiple inputs and multiple outcomes' (ibid.). In making this point, Watson decries the paucity of data on 'student learning outcomes' throughout Australia, and argues that, '...the provision of data is now essential if the Industry Commission is to establish benchmarks of system performance which reflect the effectiveness as well as the efficiency of schooling' (p. 106). However, Watson then suggests that such data '...could be collected from the range of state-mandated testing programs that are currently implemented in Australian schools' (p. 115). This has long been proposed by the gatherers and purveyors of educational performance indicators in Australia (see relevant contributions in Wyatt & Ruby, 1989; Hewton, 1990); but there are major limitations and pitfalls in doing so.

The use of educational performance indicators in the form of mean test or examination scores typically derived from such programs, tends to be very narrowly focused on a comparative *ranking* of schools rather than on identifying factors that *explain* school differences. The world-wide movement to provide such indicators (see Bottani & Tuinjman, 1994) is partly motivated by a prevailing belief that there is virtue in the mere fact of publishing comparative information. Similar sentiments

have been articulated and promulgated here in Australia (see Maslen, 1996) and evidenced in international contributions by Australian academics and senior bureaucrats (eg. Fasano, 1994; Wyatt, 1994). Indeed, since 1996, the Victorian government has promoted the publication in major daily newspapers of Year 12 student achievement data for the *Victorian Certificate of Education* – both at the individual student-level and aggregated to the school-level (see Ball, 1998; Ball & Brown, 1998; Brown, 1998). While the aggregated school-level data has been 'adjusted' for students' general 'ability' as measured by the *General Achievement Test* (see Hill, Brown, Rowe & Turner, 1997), serious problems related to interpretation and inference remain (Rowe, 1999a,b,c, 2000a,b).

For schools in the UK, the annual 'league table' rankings of schools on the basis of examination results have attracted such criticism (see Goldstein & Myers, 1996) that even the government which introduced them has conceded that they can be misleading (DfEE, 1995). Nonetheless, these 'league tables' continue to be published with wide political, community and media support. Whereas there is some consensus that appropriately contextualised, or 'value-added' comparisons are desirable (see Fitz-Gibbon, 1997; Fitz-Gibbon & Tymms, 1993; Goldstein, 1997a,b,c; Hill, 1995; Rowe, 1996; 1999a,b, 2000a,b; Saunders, 1999), they are rare and there are considerable practical difficulties in implementing them; moreover, all rankings are fallible. In fact, several studies have now shown that there are serious and *inherent* limitations to the usefulness of such indicators for providing reliable judgements about educational institutions (see Goldstein & Thomas, 1996; Goldstein & Spiegelhalter, 1996; Rowe, 1999b, 2000a,b). The reasons for these limitations are as follows:

- Given what is known about *differential school effectiveness* (eg., Nuttall, Goldstein, Prosser & Rasbash, 1989; Hill & Rowe, 1998; Rowe, 2000a; Rowe & Hill 1998), it is not possible to provide simple summaries that capture all of the important features of schools.
- By the time information from a particular school has been analysed, it refers to a 'cohort' of students who entered that school several years previously so that its usefulness for *future* students and the making of judgements about *school effectiveness* may well be dubious. Moreover, where information is analysed on a yearly basis, it is necessary to make adjustments for prior contributing factors that extend over two or more years in time. In fact, it is increasingly recognised that schools, or teachers within those schools, should not be judged by a single 'cohort' of students, but rather on their performance over time. As noted by Goldstein (1997c), this makes the historical nature of school effectiveness judgements an acute problem.
- Above all, even when suitable adjustments for students' intake characteristics and prior achievement have been taken into account, the resulting *value-added* estimates have too much *uncertainty* attached to them to provide reliable rankings. This point, illustrated in Figure 1, is vital and one that is all too-frequently ignored by advocates of published 'league tables'.

The data illustrated in Figure 1 are typical of those obtained from analyses of school-level results for a given subject area from State-wide monitoring programs

such as the New South Wales *Basic Skills Testing* (BST) program, the Victorian *Learning and Assessment Project* (LAP), the Western Australian *Monitoring Standards in Education* (MSE) program, or from Year 12 examination results such as the New South Wales *Higher Schools' Certificate* (HSC) and the *Victorian Certificate of Education* (VCE). A key feature of Figure 1 is the extent to which the 95% confidence intervals surrounding the mean point-estimates for each school cover a large part of the total range of estimates, with approximately 80 per cent of the intervals overlapping the population mean (zero). In particular, it illustrates that attempts to separate or rank schools in the form of 'league tables' are subject to considerable uncertainty.[v] Furthermore, there is always the difficulty that any statistical model used to provide such estimates will fail to incorporate **all** the appropriate adjustments, or in some other way may be mis-specified. Thus, at best, ranked 'value-added' estimates can only be used as screening devices to identify 'outliers' (which could form the basis for follow-up), but they cannot be used as definitive measures of the effect of those schools *per se* on student learning.

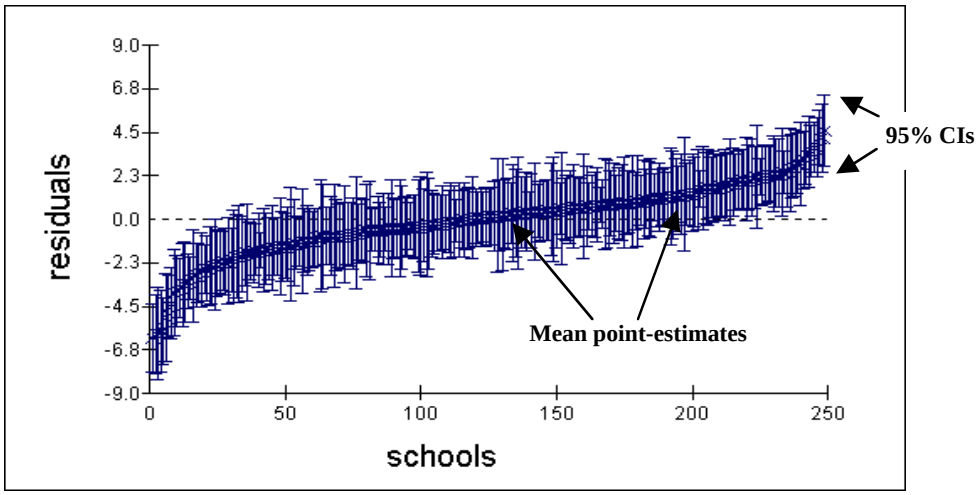


**Figure 1. Typical pattern for ranked residuals of schools' mean test/examination scores, showing schools' mean point-estimates, bounded by their 95% confidence intervals (CIs)**

It is important to emphasise that while the use of 'value-added' measures may be able to establish that differences exist among schools, in the form of 'league table' rankings they cannot, with any precision, indicate how well a particular school is performing (see Goldstein, 1997a,b,c; Saunders, 1999). The inherent uncertainty of the estimates operates as a fundamental barrier to such knowledge. It should also be stressed that raw, or even 'value-added' estimates that are ranked in this way, are ***relative*** ones; that is, they position each school in relation to other schools with which they are being compared, and at a particular point in time. If comparison groups are not representative of the population of interest, or not made on the basis

of agreed 'like-with-like' criteria, the interpretation of estimates for individual schools becomes problematic and potentially irresponsible.

Under such circumstances, the labelling of schools as 'effective' or 'ineffective' may be quite misleading unless this is understood. For those schools identified as 'effective' there is the danger of complacency on the one hand, and on the other, inordinate pressure on teachers and students to ensure that each year's examination and/or test results are an improvement on those of the previous year, without reference to student characteristics in the current cohort. It also begs the question of how 'school effectiveness' should be defined and measured. By contrast, for those schools identified as 'ineffective', experience indicates that this label frequently has the effect of lowering morale and obscures positive aspects of school functioning other than those measured by students' scores on tests and examinations. In such cases it is highly questionable whether the public humiliation of being assigned a low rank on a 'league table' is the best way to encourage remediation and improvement.

Despite these reservations, the use of adjusted school or subject level estimates to detect very discrepant units does have certain uses. For example, schemes that provide individual schools with 'value-added' feedback for their own use, such as the *A-levels Information Service* in the UK (see Fitz-Gibbon, 1992, 1996), appear to be of assistance, despite large uncertainty intervals around the estimates due to the small student numbers involved. Their use may also be worth pursuing as a device for government Departments of Education (or others) to indicate where follow-up investigations may be helpful; but this is a *higher level* monitoring function carried out on groups of schools as a screening mechanism.

However, **considerable care** is required in handling such functions and only in conjunction with other pertinent information. An important example of the lack of care is the current use of 'partially adjusted' estimates by OFSTED in England and Wales (see TES, 1996) that are based on analyses of school average scores that merely adjust for students' social background characteristics but not intake or prior achievement. In this context, Goldstein and Thomas (1996, p. 162) note: 'Used sensitively, by those charged with supporting rather than merely judging schools, such a screening procedure could have value among other sources of information'. Following a detailed discussion of the limitations endemic to the use of examination results as indicators of school performance, the concluding comments of Goldstein and Thomas (1996, p. 162) are worth noting:

> Where it is possible to study institutions over time, the time trends will provide further information which can be of interest, although relatively long time periods will be required. In addition, for A-AS-level results, the information for judgement will become available for use typically 3 years after the cohort being studied entered their institutions. In the meantime those institutions may have changed and we may need to rely on more qualitative local judgements about such changes. This makes the use of this kind of data for choice purposes, whether adjusted or not, quite problematic. It is indeed a feature of most indicator systems when used in such a way and implies that the aim in the *Parents' Charter* of using examination results for school comparisons cannot be achieved: there is no simple method of comparison which can achieve fair and accurate comparisons between institutions. It follows that the

publication of league tables without a clear statement of their limitations is both misleading and scientifically unjustifiable. This point is also made in a report by the School Curriculum and Assessment Authority on value-added measures (SCAA, 1994, chapter 3, section 3b).

For all these reasons, performance indicators based on the ranking of schools' average examination and test scores have little to offer in shedding light on *school effectiveness*. This is not, of course, to deny that individual schools should be held accountable through the collection of relevant performance information. Rather, it is suggested that it is highly unsatisfactory to attempt this, *principally and indirectly*, by invoking 'evidence' based on the achievement scores of their students.

## Towards responsible use of educational performance information

Throughout industrialised societies there is a prevailing strong belief that the publication of information about the performance of public bodies is an overwhelming social good. In some societies, such as the United States, it is enshrined in public disclosure legislation. In the context of 'school effectiveness' the role of published performance information is crucial. Whether intended or not, it provides the data base for comparative judgements, or in market terms, it introduces a common currency by which the relative 'worth' of schools are measured. Indeed, this appears to be the primary purpose of such information, and political discourse implicitly acknowledges this whenever reference is made to such matters as 'parental choice' or 'raising educational standards'.

As a reaction to unreasonable secrecy, the belief in open access to information seems wholly healthy and has led to many benefits. Nevertheless, it can be argued strongly that the public disclosure of information cannot be held to be an *absolute* principle. This is recognised by governments, for example, who normally reserve the right to withhold information deemed to threaten the 'security' of a nation. Similarly, if the publication of certain information has the potential for harming individuals, or may be seriously misleading, then a justifiable case can be mounted for refusing its publication. It could be contended that much of what might be described as *educational performance indicators* based on measures of student achievement falls into this category. Its ability to reflect objective reality may be extremely limited, and its publication may therefore cause both misleading and incorrect inferences about schools and 'school effectiveness' to be drawn.

In such circumstances, there is strong case for withholding publication. If for whatever reason, publication cannot be prevented then the information should have appropriate warnings attached about its interpretation. By this is meant not simply warnings of the kind that appears on tobacco advertisements, but a proper and prominent *explanation* of why the information is suspect, together with an assurance that the publishers of the information are fully aware of and accept its limitations.

This view invites criticism of much of the activity that comes under the rubric of *educational performance indicators*. A great deal of this information is produced merely because the data happen to be available (eg., Bottani, 1994). Some

of it, such as the achievement scores derived from international studies of mathematics and science (Rotberg, 1990), have been taken, even usurped, by governments and by international agencies such as OECD in order to rank countries in a supposed 'order of merit'. Even where relevant caveats are included in published reports, they tend to be of little avail, and the overall message is that the information presented is useful and informative.

In spite of the these problems, accountability pressures on governments are not likely to abate in the foreseeable future; nor is the demand for published educational performance indicators based on students' test and examination results obtained from large-scale monitoring programs likely to diminish. Given this 'reality', it is very much in the interests of those wishing to publish such information to consider carefully the need to provide proper guidelines for their publication, if for no other reason than to minimise the risk of widespread public distrust in the face of manifestly poor and misleading information, and to avoid a possible wholesale rejection of all information about schools and schooling – both good and bad. To the writer's credit, Watson (1996, p. 120) recognises the need for such guidelines by proposing three 'principles' that '...should underpin any performance indicators framework', namely: (1) the need to develop multiple outcomes, '...which reflect the wide spectrum of objectives for education, not just cognitive outcomes' (ibid.), (2) the need to account for contextualisation factors and to ensure that only 'like-with-like' comparisons are made, and (3) the need for published reports to convey '...the limitations of performance indicators for policy decisions' (ibid.). Watsons' principles constitute a useful start, but given the complexities endemic to the issues involved, a more detailed elaboration is required, particularly for issues related to *publication*.

In an attempt to provide a relevant set of *publication standards*, Goldstein and Myers (1996) and Myers and Goldstein (1996) propose a set of basic principles for what they refer to as *a code of ethics for performance indicators*; they state:

> Just as educational test constructors have ethical guidelines and in most societies there are codes governing the publication of pornographic or derogatory materials, so we believe there should be a code for the publication of comparative institutional information. ... Our aim is to start a public discussion to see if some consensus can be reached about what a suitable code might contain and whether and how it might be enforced (p. 4).

In setting out these guidelines, Goldstein and Myers consider the various users of performance indicator information. For example, they suggest that policy makers are interested in broad questions of 'effectiveness' whereas parents and students tend to be more concerned with local details relevant to their particular needs. For all users, however, there is a shared interest in accuracy and general quality and it is these factors which motivate two basic principles:

***1.*** ***The principle of unwarranted harm***. The fundamental guiding principle, as with many ethical codes, is that the publication or communication by other means, should cause no *unwarranted* harm to those who are identified. The term *unwarranted* is used since there will clearly be legitimate circumstances

when it is in the 'public interest' for genuinely poor performance to be made known. Nevertheless, the principle is that innocents should be protected from misleading insinuations: for example, implying that a ranking of schools by test or examination scores is also a ranking of educational 'quality' or 'merit'.

2. ***The principle of the right to information***. Given that the information available is believed to accurate and relevant, there shall be a presumption that it be made public, but modified by the first principle where necessary.

These two principles require some elaboration to be applied in practice. The following points can be viewed as offering guidance on the application of principles 1 and 2:

- ***Contextualisation***. Indicators should provide information that allow for fair comparisons. Indicators strongly affected by extrinsic/contextual factors (such as student intake characteristics) should not be used unless adjustments have been made for those characteristics. For example, school rankings based solely on 'raw' examination or test score results should not be published. All adjustments for contextual factors should be described carefully and displayed prominently.
- ***Presentation of uncertainty***. All performance indicators should be accompanied by estimates of statistical uncertainty such as those illustrated in Figure 1. These should reflect sampling variability, and where possible, the uncertainty due to choice of measurement, statistical techniques used, and so on. The presentation of uncertainty intervals shall be as prominent as those for the indicator values themselves.
- ***Multiple indicators***. Where possible, multiple indicators relevant to each institution should be presented, rather than a single or summary one. This should be done to avoid undue concentration on any one aspect of performance.
- ***Institutional response***. Any institution for which there is a set of published indicators shall have the right to question the accuracy of information about it. Compilers of indicators shall be obliged to make data available in a format which allows for re-analysis of those data by a responsible and competent 'third party', subject to appropriate confidentiality constraints and guided by principle 1.
- ***Agency responsibilities***. Agencies responsible for providing public performance indicators shall assume a responsibility for disseminating accurate and informative material about the underlying procedures used for compilation. They should make relevant technical information accessible, including details of the sampling and statistical methods of analysis used. There is also a responsibility for secondary providers such as the media (newspapers, radio, television) to inform the public of the strengths and limitations of the indicators.
- ***Enforcement***. One would hope that the process of developing such guidelines would generate sufficient awareness of their importance and a common interest in abiding by them. Nevertheless, it may be necessary to establish formal

> regulatory mechanisms to ensure compliance. This is clearly a matter for careful consideration, but a start in this country might be made with the involvement of professional bodies such as the Australian Association for Research in Education (AARE) and the various Boards of Studies in each State, or with an organisation like the Australian Council for Educational Research (ACER). Ultimately, the appointment of an educational ombudsperson could provide a means of appropriate redress for aggrieved persons and/or institutions (schools).

In setting out these principles for consideration, the intention of Goldstein and Myers (1996) and Myers and Goldstein (1996) is to challenge conventional assumptions about the publication of educational performance information, and to highlight the complexity that surrounds these issues. As with any code of ethics, a primary function is to raise awareness of the problems and benefits resulting from particular courses of action. What is important is that persons and institutions (schools) should have a means of redress if there is cause to believe they are being unfairly labelled. Moreover, those who are exposed to the information should also be exposed to views about its limitations, as well as to its prima facie justification. Governments and their bureaucracies have a special responsibility here. Despite prevailing cynicism about officialdom, it is nevertheless the case that the mere fact of publishing information by an official body lends it credence. It is therefore important that the publication makes every attempt at honesty and accuracy, since after all, it is the fundamental responsibility of those privileged with access to information and the means to process it, to present it fairly.

## Concluding Comments

Behind the publication of educational performance indicators in the form of 'league tables' lie unspoken assumptions and value judgements about the location of 'blame' or 'credit'. As already suggested, the underlying assumption is that if a school is deemed to be 'effective' or 'ineffective' in terms of the ranked position of its students' average test or examination scores on a 'league table', the reason for that performance resides in the school. Even the contextualisation of performance using adjusted or 'value-added' scores may strengthen such an assumption by encouraging the view that **all** other factors have been accounted for, so that any residual variation has its origin in the school. The inherent imprecision of all performance measures and the provisional nature of any conclusions, as argued here, needs to be stressed. Indeed, Saunders (1999, pp. 253-254) expresses a relevant warning in the following terms:

> …both researchers and policy-makers…have a duty to be clear about the fact that there are value judgements as well as conceptual assumptions and technical decisions implicit in what they choose to measure; and that 'value added' measures of effectiveness – powerful as they often are for analytical purposes – are dependent for their credibility on the degree to which those judgements are publicly articulated.

The issue of contextualisation in school education is an important one, but it extends well beyond simplistic notions of 'value-added' indicators of performance. At the very least it needs to be extended to include the general political and social

context within which schools operate. Education is not a one way enterprise. It is not simply the case that the performance of persons in the workplace or society at large can be causally related directly to their education. To attribute the poor economic performance of a nation to the performance of its education system, for example, is to make both a logical and empirical blunder. It is just as easy to argue the reverse, namely that the economic performance of a nation has direct effects on its education system in terms of motivation, resource provision (eg., Raffe & Willms, 1991), and other crucial input mechanisms such as the quality and quantity of teacher professional development (Hill & Crévola, 1999; Rowe, 1996; Rowe & Hill, 1998; Rowe & Rowe, 1999; Rowe & Sykes, 1989). Certainly it is not legitimate to argue, as is frequently the case, that 'league tables' of international educational performance reflect the quality of national educational systems. The attribution of cause and effect is replete with difficulties in such circumstances, and the mere repetition of any given interpretation does not strengthen its plausibility. The same logic applies to the growing use of 'league tables' of educational achievement data to judge the 'quality' or relative 'worth' of individual schools.

The existence of an accountability climate that insists on providing published information which invites comparative judgements about the relative 'worth' of schools – and, inevitably, about the teachers who work in them – is problematic. It is a social and political minefield that has the potential for considerable harm unless it is handled with great care. Again, this is not to deny the usefulness of school-level educational performance indicators involving student achievement data, provided that relevant contextual factors have been taken into account and that the statistical uncertainty associated with the estimates obtained are displayed prominently. McGaw (1991, p. 138) points out the benefits and risks involved in universal achievement monitoring programs in the following terms:

> The benefit of assessing all students is that each school obtains information about its program and teachers obtain potentially helpful diagnostic information about all students. The risk is that the universality of such a program will allow and even encourage comparisons among schools, without consideration of the effect of non-school factors on scores, and so oblige schools to concentrate more upon specific preparation for the tests.

While it would be preferable to implement assessment programs at the beginning of the school year solely for *diagnostic* purposes to assist teachers, as in France (see OECD, 1993), accountability pressures on State and Federal governments in Australia to monitor educational standards are political realities, and ones that are likely to increase. In one sense it could be argued that to propose a control mechanism in the form of a *code of ethics* for the publication of educational performance indicators of the kind outlined above is akin to 'throwing a wet fish at a runaway train'. But if we, as a society, do nothing, we run the grave risk of rejecting the good and useful information because it cannot be distinguished from the bad and misleading. That, to put it mildly, would be a disaster. An even greater disaster would be, that in our efforts to meet increasing demands for *assessment, accountability, performance indicators, standards monitoring, quality assurance, school effectiveness* and (now) *benchmarking,* we lose sight of ensuring that what we offer in school education is accessible to **all** students. 'The provision of

universal education was one of the great social and moral triumphs of the modern period. Universal **success** should be the aim of the postmodern' (Wilson, 1996, p. 8). We stand forever condemned if seduced into diverting the focus of our efforts elsewhere. 'Let's get real'!

## References

Ball, S. (1998) Putting value into the VCE. *The Age*, December 16, 1998, VCE Supplement: How your school performed, p. 3.

Ball, S., & Brown, T. (1998) Part of the big picture. *Herald Sun*, December 16, 1998, VCE Supplement: How they fared, p. 41.

Berlak, H. (1992) The need for a new science of assessment. In H. Berlak, F. Newmann, E. Adams, D. Archbald, T. Burgess, J. Raven, & T. Romberg (Eds.) *Toward a new science of educational testing and assessment*. Albany: State University of New York Press.

Board of Studies (1994) *Curriculum and Standards Framework.* Melbourne, Vic: Board of Studies.

Board of Studies (1999) *Cuuriculum & Standards Framework II*: *Draft for Consultation*. Melbourne, Vic: Board of Studies.

Bosker, R., Kremers, E., & Lugthart, E. (1990). School and instructional effects on mathematics achievement. *School Effectiveness and School Improvement*, Vol. 1, pp. 213-248.

Bosker, R., & Scheerens, J. (1989). Issues in the interpretation of the results of school effectiveness research. *International Journal of Educational Research*, Vol. 13, p. 741-751.

Bosker, R. & Scheerens, J. (1994). Alternative models of school effectiveness put to the test. *International Journal of Educational Research*, Vol. 21, No. 2, pp. 159-180.

Bottani, N. (1994) The OECD international education indicators. *Assessment in Education*, Vol. 1, pp. 333-350.

Bottani, N. & Tuijnman, A. (1994) The design of indicator systems. In A.C. Tuijnman, & T.N. Postlethwaite, T.N. (Eds.) (1994) *Monitoring the standards of education: Papers in honour of John P. Keeves.* Oxford: Pergamon.

Broadfoot, P. (1994) The myth of measurement. *Inaugural Address*, The University of Bristol.

Broadfoot, P. (1996) Assessment and learning: Power or partnership? In H. Goldstein and T. Lewis (Eds.) *Assessment: Problems, developments and statistical issues.* Chichester: John Wiley & Sons Ltd.

Brown, T. (1998) Measure for measure, how the figures work. *The Age*, December 16, 1998, VCE Supplement: How your school performed, p. 2.

Cannell, J. (1988) Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice*, Vol. 7, 5-9.

Chapman, J., Angus, L., Burke, G., & Wilkinson, V. (Eds.) (1991) *Improving the quality of Australian schools*. Australian Education Review No. 33. Hawthorn, Vic: The Australian Council for Educational Research.

Cuttance, P. (1990) Performance indicators and the management of quality in education. In J. Hewton (Ed.) *Performance indicators in education: What can they tell us?* Brisbane: Australian Conference of Directors-General of Education.

Darling-Hammond, L. (1994) Performanced-based assessment and educational equity. *Harvard Educational Review*, Vol. 64, pp. 5-29.

Dawkins, J. (1988) *Strengthening Australia's schools: A consideration of the focus on the content of schooling*. Canberra, ACT: Australian Government Publishing Service.

DES (1991) *The Parents' Charter*. Department of Education and Science, London: Her Majesty's Stationery Office.

DfEE (1995) *GCSE to GCE A-AS value added: Breifing for schools and colleges.* London: Department for Educational and Employment.

DfEE (1996a) *Results of the 1995 National Curriculum Assessments of 7 Year Olds in England*. London: Department for Education and Employment.

DfEE (1996b) *Results of the 1995 National Curriculum Assessments of 11 Year Olds in England*. London: Department for Education and Employment.

DfEE (1996c) *Results of the 1995 National Curriculum Assessments of 14 Year Olds in England*. London: Department for Education and Employment.

Donnely, K. (1999) An international comparative analysis across educational systems: Benchmarking the Victorian Curriculum and Standards Framework. *IARTV Seminar Series*, No. 83, May 1999.

Fasano, C. (1994) Knowledge, ignorance and epistemic utility: Issues in the construction of indicator systems. In *Making education count: Developing and using international indicators.* Paris: OECD.

Fitz-Gibbon, C.T. (1992) School effects at A level: Genesis of an information system? In D. Reynolds and P. Cuttance (Eds.) *School effectiveness: Research policy and practice*. London: Castle.

Fitz-Gibbon, C. (1996) *Monitoring education: Indicators, quality and effectiveness*. London: Castle.

Fitz-Gibbon, C. (1997) *The value added national project: Final report. Feasibility studies for a national system of valued added indicators.* London: SCAA.

Fitz-Gibbon, C., & Tymms, P. (1993) *Value added: A perspective on the contribution from examination results* (CSCS Discussion Paper 1) Northhampton (UK): Centre for the Study of Comprehensive Schools.

Floden, R. (1994) Reshaping assessment concepts. *Educational Researcher,* Vol. 23, No. 2, p. 4.

Gipps, C. (1996) *Assessment for the millennium: Form, function and feedback.* Inaugural professorial lecture delivered at the Institute of Education, University of London on June 6, 1996. Institute of Education, University of London.

Gipps, C. & Murphy, P. (1994) *A fair test? Assessment, achievement and equity*. Buckingham, Open University Press

Goldstein, H. (1997a, March) Value added tables: The less-than-holy-grail. *Managing Schools Today*, pp.18-19.

Goldstein, H. (1997b, July 18) From raw to half-baked. *Times Educational Supplement*, p. 15.

Goldstein, H. (1997c) Methods in school effectiveness research. *School Effectiveness and School Improvement,* Vol 8, pp. 369-395.

Goldstein, H. & Cuttance, P. (1988) A note on national assessment and school comparisons. *Journal of Education Policy,* Vol. 3, pp. 197-202.

Goldstein, H. & Healy, M. (1995) The graphical presentation of a collection of means. *Journal of the Royal Statistical Society,* A, Vol. 158, pp. 175-177.

Goldstein, H. & Lewis, T. (Eds.) (1996) *Assessment: Problems, developments and statistical issues*. Chichester: John Wiley & Sons Ltd.

Goldstein, H. & Myers, K. (1996) Freedom of information: Towards a code of ethics for performance indicators. University of London Institute of Education.

Goldstein, H. & Spiegelhalter, D. (1996) League tables and their limitations: Statistical issues in comparisons of institutional performance. With discussion. *Journal of the Royal Statistical Society,* A, Vol. 159, pp. 385-443.

Goldstein, H. & Thomas, S. (1996) Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society*, A, Vol. 159, pp. 149-163.

Hannan, W. (1995) Why teach to outcomes? *IARTV Occasional Paper*, No. 40, May, 1995.

Harlen, W. (Ed.) (1994) *Enhancing quality in assessment*. London: Paul Chapman Publishers Ltd.

Hewton, J. (Ed.) (1990) *Performance indicators in education: What can they tell us?* Brisbane: Australian Conference of Directors-General of Education.

Hill, P. (1994) Putting the national profiles to use. *Unicorn,* Vol. 20, No. 2, pp. 36-42.

Hill, P. (1995) Value added measures of achievement. *IARTV Seminar Series*, No. 44, May, 1995.

Hill, P., Brown, T., Rowe, K.J., & Turner, R. (1997) Establishing comparability of Year 12 school-based assessments. *Australian Journal of Education,* Vol. 41, No. 1, pp. 27-47.

Hill, P. & Crévola, C. (1999) The role of standards in educational reform for the 21st century. In D. Marsh (Ed.), *ASCD Year Book 1999: Preparing our schools for the 21st century*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hill, P. & Rowe, K.J. (1996) Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement,* Vol. 7, No. 1, pp. 1-34.

Hill, P. & Rowe, K.J. (1998) Modelling student progress in studies of educational effectiveness. *School Effectiveness and School Improvement,* Vol. 9, No. 3, pp. 310-333.

Istance, D. & Lowe, J. (1991) Schools and quality: The concept and the concern. In J. Chapman, L. Angus, G. Burke, & V. Wilkinson (Eds.) (1991) *Improving the quality of Australian schools.* Australian Education Review No. 33, pp. 22-49. Hawthorn, Vic: The Australian Council for Educational Research.

Kellaghan, T., Madaus, G., & Airasian, P. (1992) *The effects of standardized testing*. Boston: Klewer-Nijhoff Publishing.

Lacey, C. & Lawton, D. (1981) *Issues in evaluation and accountability*. London: Methuen.

Lokan, J. & Ford, P. (1994) *Mapping state testing programs.* Report prepared for the National Industry Education Forum (NIEF) Melbourne, Vic: The Australian Council for Educational Research.

Marginson, S. (1993) *Education and public policy in Australia*. Cambridge: Cambridge University Press.

Marsh, D. (Ed.) (1999) *ASCD Year Book 1999: Preparing our schools for the 21st century*. Alexandria, VA: Association for Supervision and Curriculum Development.

Maslen, G. (1996) League tables provoke uproar. *Times Education Supplement*, September 27, 1996, p. 18.

Masters, G. (1990) Improving the assessment of student outcomes. In J. Hewton (Ed.), *Performance indicators in education: What can they tell us?* Brisbane: Australian Conference of Directors-General of Education.

Masters, G. (1991) *Assessing Achievement in Australian Schools*: A discussion paper commissioned by the Industry Education Forum. Hawthorn, Vic: The Australian Council for Educational Research.

Masters, G. (1994) *Setting and measuring performance standards for student achievement*. Paper presented at the conference 'Public Investment in School Education: Costs and Outcomes', sponsored by The Schools Council and The Centre for Economic Policy Research, Australian National University, Canberra, March 17, 1994.

McGaw, B. (1991) Monitoring education systems. In J. Chapman, L. Angus, G. Burke, & V. Wilkinson (Eds.) (1991) *Improving the quality of Australian schools.* Australian Education Review No. 33 (pp. 134-139) Hawthorn, Vic: The Australian Council for Educational Research.

Mortimore, P. (1991) School effectiveness research: Which way at the crossroads? *School Effectiveness and School Improvement*, Vol. 2, No. 3, pp. 213-229.

Mortimore, P. (1998) *The road to improvement: Reflections on school effectiveness.* Lisse, The Netherlands: Swetz & Zeitlinger.

Moss, P. (1994) Can there be validity without reliability? *Educational Researcher*, Vol 23, pp. 5-12.

Murphy, R. & Broadfoot, P. (Eds.) (1995) *Effective assessment and the improvement of education.* London: The Falmer Press.

Myers, K. & Goldstein, H. (1996) Failing schools in a failing system. *Association for Supervision and Curriculum Development (ASCD) Year Book*.

National Centre on Education and the Economy (1997) *New standards for performance standards*. Washington, DC: Author.

Newmann, F. & Archbald, D. (1990) Organizational performance of schools. In P. Reyes (Ed.) *Teachers and their workplace: Commitment, performance and productivity*. Newbury Park, CA: Sage Publications.

Nisbet, J. (1993) Introduction. In *OECD - Curriculum reform: Assessment in question*. Paris: Organisation for Economic Cooperation and Development.

Nuttall, D., Goldstein, H., Prosser, R., & Rasbash, J. (1989) Differential school effectiveness. *International Journal of Educational Research,* Vol. 13, No. 7, pp 769-776.

OECD (1983) *Compulsory schooling in a changing world.* Paris: Organisation for Economic Cooperation and Development.

OECD (1986) *Education and training for manpower development.* Paris: Organisation for Economic Cooperation and Development.

OECD (1989) *Schools and quality: An international report*. Paris: Organisation for Economic Cooperation and Development.

OECD (1993) *Curriculum reform: Assessment in question*. Paris: Organisation for Economic Cooperation and Development.

Raffe, D. & Willms, J. (1991) Schooling the discouraged worker: Local labour-market effects on educational participation. *Sociology,* Vol 23, p29 559-581.

Hill, P.Hill, P.Vol. 41, No. 1, pp.Hill, (1999) Empirical validity for a comprehensive model on educational effectiveness. *School Effectiveness and School Improvement,* P. & Crévola 193-216.

Rowe, K.J. (1996a) Assessment, performance indicators, league tables, value added measures and school effectiveness: Issues and implications. *IARTV Seminar Series*, No. 58, October, 1996.

Rowe, K.J. (1996b) Lay off teachers! *Education Guardian*, February 6, 1996, p. 8.

Rowe, K.J. (1999a) *Assessment, performance indicators, 'league tables', 'value-added' measures and school effectiveness? Consider the issues and 'let's get real'*! Paper presented at the 1999 AARE-NZARE Joint Conference of the

Australian and New Zealand Associations for Research in Education, Melbourne Convention Centre, November 29 – December 2, 1999 (Index Code: ROW99656)

Rowe, K.J. (1999b) *VCE Data Project (1994-1998): Concepts, issues, directions & specifications*. Melbourne, Vic: Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J. (1999c) *Reliability of assessments for VCE studies: 1995-1998*. Melbourne, Vic: Centre for Applied Educational Research, The University of Melbourne.

Rowe, K.J. (2000a) Educational performance indicators. In M. Forster, G. Masters and K.J. Rowe, *Measuring learning outcomes: Options and challenges in evaluation and performance monitoring* (pp. 2-20) Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring. Washington, DC: The World Bank Institute.

Rowe, K.J. (2000b) *Multilevel structural equation modeling with MLn/MLwiN & LISREL8.30: An integrated course* (4th ed.) The 7th ACSPRI Winter Program in Social Research Methods and Research Technology, The University of Queensland. Camberwell, Vic: The Australian Council for Educational Research [ISBN 0 86431 357 8].

Rowe, K.J., & Hill, P.W. (1996). Assessing, recording and reporting students' educational progress: The case for 'Subject Profiles'. *Assessment in Education,* Vol. 3, pp. 309-352.

Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation,* Vol. 4, No. 4, pp. 307-347.

Rowe, K.J., & Hill, P.W., & Holmes-Smith, P. (1995). Methodological issues in educational performance and school effectiveness research: A discussion with worked examples. *Australian Journal of Education,* Vol. 39, pp. 217-248.

Rowe, K.J., & Rowe, K.S. (1999). Investigating the relationship between students' *attentive-inattentive* behaviors in the classroom and their literacy progress. *International Journal of Educational Research*, Vol. 31, Nos. 1-2 – Whole Issue, pp. 1-138.

Rowe, K.J. & Sykes, J. (1989) The impact of teacher professional development on teachers' self-perceptions. *Teaching and Teacher Education,* Vol. 5, pp. 129-141.

Rowe, K.J., Turner, R., & Lane, K. (2000) Performance feedback to schools of students' Year 12 assessments: The *VCE Data Project.* In R. Coe and A.

Visscher (Eds.) *School improvement through performance feedback*. Lisse, The Netherlands: Swetz & Zeitlinger.

Rotberg, I. (1990) I never promised you first place. *Phi Delta Kappan*, December, 1990, pp. 296-303.

Sanders, W. & Horn, S. (1994) The Tennessee value-added assessment system (TVASS): Mixed model methodology in educational assessment. *Journal of Personnel Evaluation in Education*, Vol. 8, pp. 299-311.

Saunders, L. (1999) A brief history of educational 'value added': How did we get to where we are? *School Effectiveness and School Improvement*, Vol. 10, No. 2, pp. 233-256.

Scheerens, J. (1993) Basic school effectiveness research: Items for a research agenda. *School Effectiveness and School Improvement*, Vol. 4, pp. 17-36.

Scheerens, J. (1995) *School effectiveness as a research discipline: Some comments on country reports*. Paper presented at the eighth International Congress for School Effectiveness and Improvement, CHN, Leeuwarden, The Netherlands, January 3-6, 1995.

Scheerens, J., & Bosker, R. (1997) *The foundations of educational effectiveness*. Oxford: Pergamon.

Shavelson, R.J. (Guest Editor) (1994) Performance assessment. *International Journal of Educational Research*, Vol. 21, pp. 233-350.

Smith, M.L. (1991) Put to the test: The effects of external testing on teachers. *Educational Researcher*, Vol. 20, pp. 8-11.

Smith, M. & Rottenburg, C. (1991) Unintended consequences of external testing in elementary schools. *Educational Measurement: Issues and Practice*, Vol. 10, pp. 7-11.

Taylor, C. (1994) Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal*, Vol. 31, pp. 231-262.

TES (1996) Inspectors take account of deprivation. *Times Educational Supplement*, February 23, 1996.

Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997) Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years (Leading artcile) *School Effectiveness and School Improvement*, Vol. 8, pp. 169-197.

Tucker, M. & Codding, J. (1998) *Standards for our schools: How to set them, measure them and reach them*. San Fransico, CA: Jossey-Bass.

Tuijnman, A. & Postlethwaite, T. (Eds.) (1994) *Monitoring the standards of education: Papers in honor of John P. Keeves*. Oxford: Pergamon.

Tymms, P. (1993) Accountability – can it be fair? *Oxford Review of Education*, Vol. 19, pp. 291-299.

Watson, L. (1996) Public accountability or fiscal control? Benchmarks of performance in Australian schooling. *Australian Journal of Education*, Vol. 40, pp. 104-123.

Wiggins, G. (1989) Teaching to the (authentic) test. *Educational Leadership*, Vol. 46, pp. 41-47

Wilson, B. (1993) National curriculum and national profiles: Development, implementation and use. *IARTV Seminar Series*, No. 23, April, 1993.

Wilson, B. (1996) Current educational priorities, future directions and initiatives. *IARTV Occasional Paper*, No. 45, May, 1996.

Wood, R. (1986) The agenda for educational measurement. In D.L. Nuttall (Ed.) *Assessing educational achievement*. London: The Falmer Press.

Woodhead, C. (1996) *The annual report of Her Majesty's Chief Inspector of Schools: Standards and quality in education 1994/95*. Office for Standards in Education (OfSTED) London: HMSO.

Wyatt, T. (1994) Education indicators: A review of the literature. In *Making education count: Developing and using international indicators*. Paris: OECD.

Wyatt, T. & Ruby, A. (Eds.) (1989) *Education indicators for quality, accountability and better practice: Papers from the second national conference, Surfers Paradise, 1989*. Sydney: Australian Conference of Directors-General of Education.

## ENDNOTES

[i] These were published in the form of three booklets (DfEE, 1996a,b,c), one for each of the three age groups, and released to British members of parliament (MP's) and media on January 26, 1996. For several days following, television, radio and all national daily news-papers, including the London tabloids, were replete with 'front page' reports, 'editorials' and 'leading articles' commenting on the "poor" results (especially for 11 year olds), mostly in the context of heated exchanges among government and opposition MP's in the House of Commons. Initially, much of the reported blame for the "poor" results was generated from the point-scoring antics of MP's accusing each other, *inter alia*, of political and administrative "incompetence". However, upon the release of the OfSTED report (Woodhead, 1996) a few days later, there was a savage shift in blame to teachers, and primary teachers in particular (see Note 2)

[ii] On February 6 1996, the British media had a 'field day' with selected portions of *The Annual Report of Her Majesty's Chief Inspector of Schools: Standards and Quality in Education 1994/95* (Woodhead, 1996), from the Office for Standards in Education (OfSTED) *The Times* editorial on that day (p. 19) was entitled "*Woodhead paints a gloomy picture of teaching practice*" and referred to "...the dip in standards in the second half of primary schooling.." being due to "...the educational philosophy and poor quality of many teachers". "Woodhead identifies about 15,000 teachers who can be classed as 'poor' - some may be irredeemably so, and should not remain in their jobs". Similar headlines and sub-editor by-lines were not short on rhetoric, for example: "Half of schools failing their pupils" (*The Independent*, p.1) - cf. "Nearly half the schools in England are failing their pupils" (*The Guardian*, p.1); "Inspector condemns 25 years of weak and trendy teaching" (The Daily Telegraph, p. 8); and "Black mark for half of primaries: Teaching is blamed for low standards" (*The Times*, p. 1) Throughout this fulmination, the quality of the arguments made on the limited evidence gathered and provided was so poor (in the opinion of the present author) that a response letter to *The Guardian* was written, which was subsequently published (Rowe, 1996b).

[iii] In the Australian context, a rationale for educational monitoring is provided by McGaw (1991). In discussing the limitations of educational performance indicators and their potential misuse, Watson (1996) suggests an approach to measuring system performance designed to establish 'benchmarks' that reflect both the "efficiency and effectiveness of Australian school systems". A detailed review of current monitoring programs throughout Australia is provided by Lokan and Ford (1994).

[iv] The publication of UK-type 'league tables' of school-level public examination results in Australian states, is a relatively recent phenomenon. In Victoria, for example, such crude information was first published in major daily newspapers in December, 1996. While some modifications have since been made to its presentation (see Ball, 1998; Ball & Brown, 1998), in the view of the present author, the information remains misleading and hence, irresponsible – for the reasons outlined following.

[v] 95% Confidence Intervals for a statistic (a mean point-estimate for each school in this case – $\bar{X}_s$) are calculated from: $\bar{X}_s \pm 1.96 \times$ the school's standard error (i.e., the school's standard deviation divided by the square root of the school sample size – $\sigma_s / \sqrt{n_s}$) These intervals imply that we can be confident within 95% limits that the estimate of a school's mean lies between these upper and lower limits. However, in the present context of making comparative judgements about the relative performance of schools, these limits are more properly referred to as *uncertainly intervals*. That is, when the confidence intervals for two or more schools overlap, there is *no certainty* that their relative performance differs in any way – least of all statistically significantly. For a presentation and discussion of the relevant conceptual and technical issues, see Goldstein and Healy (1995); Goldstein and Spiegelhalter (1996); Rowe, Turner and Lane (2000).