public administration

doi: 10.1111/j.1467-9299.2010.01812.x

TOWARDS STANDARDS FOR THE PUBLICATION OF PERFORMANCE INDICATORS IN THE PUBLIC SECTOR: THE CASE OF SCHOOLS

SJOERD KARSTEN, ADRIE J. VISSCHER, ANNE BERT DIJKSTRA AND RENÉ VEENSTRA

Since the 1990s, the media and public authorities in many developed countries have published data on the quality of public services such as schools, hospitals and police. In the education sector the publication of performance indicators and league tables generated considerable debate. In this article, the policy context of this development in the education sector is analysed in conjunction with the problems associated with the public reporting of school performance indicators. This is followed by a discussion of the need for an accepted set of publication standards. The aim of this article is to initiate a debate about whether international consensus can be reached on the content of such a set of standards, and whether a particular set of standards, developed in the Dutch context, is applicable in and relevant to other countries. These Dutch standards have been applied to a variety of English, American and Dutch publications. The results of the analysis suggest that if the Dutch standards were applied, school performance publications could be significantly improved.

INTRODUCTION

The publication of reports which reflect the quality of individual schools, such as league tables or report cards, has become a prominent feature of the education system in many countries. This systematic publication began in the United States in the 1980s, spread to England and France in the early 1990s, and has since been adopted by many other developed nations (West and Pennell 2000; Karsten *et al*. 2001). In the US, in January 2002, the then president, President George W. Bush, signed into law the No Child Left Behind Act NCLB of 2001, which requires all US states to test students regularly and to hold schools and districts accountable for student performance. The ratification of NCLB followed a decade of attempts across the US to use high-stakes accountability to drive school improvement. The new law requires that the National Assessment of Educational Progress (NAEP) be administered biennially, in math and reading, to a sample of fourth and eighth grade students in each state, thereby providing a standard by which state judgements about proficiency may be compared. Furthermore, the law imposes a series of corrective actions on schools and districts which fail to demonstrate 'adequate yearly progress'. The ratification of the NCLB legislation represents an example of the influence of the world-wide accountability-based reform movement, which stresses the importance of the publication of school performance indicators for school improvement (Peterson and West 2003).

School performance publications are primarily intended to provide parents and tax-payers with information on the quality of individual schools. By including information on student outcomes, the publications can also serve as a policy tool in generating pressure for school improvement (Ladd and Walsh 2002). It should be noted here that although

Sjoerd Karsten is in the Faculty of Social and Behavioural Sciences, University of Amsterdam. Adrie J. Visscher is in the Department of Educational Organization and Management, University of Twente, Anne Bert Dijkstra is in The Netherlands' Education Inspectorate, Utrecht. René Veenstra is in the Department of Sociology, University of Groningen.

Public Administration Vol. 88, No. 1, 2010 (90–112)
© 2010 Blackwell Publishing Ltd

average student achievement levels per school are not the only published indicators of school quality – others include teacher quality and student drop-out rates – they are considered the most important. Although there is a strong belief in many countries that the publication of information about the functioning of schools, and other public institutions, serves the common good, the introduction of report cards has generated considerable public debate. Much of the debate has concentrated on the imprecision of the indicators published (see, for example, Goldstein and Thomas 1995; Visscher 2001), and on the possible negative side effects of these publications on the functioning of schools, for example, discouraging teaching staff and teaching to the test (see Foxman 1997; Karsten *et al*. 2001). Some authors criticize the public release of performance data as a neo-liberal project for introducing the market model into the public sector – and which has, therefore, to be repudiated (Apple 2001). Others argue that new regimes of public accountability have been developed to such an extent that accountability is no longer simply one component of the education system but 'constitutes the system itself' (Ranson 2003). Although we recognize that our approach is not neutral, we believe that enough is known now about the construction and effects of public performance data to allow for a substantial modification of their use and presentation.

In an overview of the problems associated with publishing school performance indicators, Visscher (2001) cited several reasons why it is often impossible for a school report card to capture an accurate picture of a school's performance. One reason given was that indicators may often be computed on the basis of a limited number of observations. In Europe, for example, indicators may be based on the exam scores of only those students who sit central examinations for optional subjects. In the US on the other hand, states are required to assess students in each of the grades 3–8, but only in one grade at the high school level. Another reason is that it is not feasible to adjust indicator computations for all relevant factors, for example, student drop-out and mobility rates among schools.

Moreover, cross-sectional indicators reflect how schools have performed, relative to other schools, in a given school year. This means that, theoretically, even the best school may have performed relatively poorly. Conversely, if that standard were very high, a relatively weak school may actually have performed quite well, in absolute terms. In recognition of this problem, in recent years, state accountability systems in the US have de-emphasized annual performance snapshots in favour of indicators which report the progress schools have made toward challenging, long-term goals.

Although there is broad acknowledgement of the significant problems associated with the publication of school performance indicators, these indicators continue to be published widely and, in the US under NCLB, have become even more prevalent. The NCLB legislation, with all its requirements of public benchmarking, provides many more parents with more detailed information about what children are learning at their child's school. At least two consequences may result from this new transparency. Firstly, parents may express concerns about the performance of their child's school directly to school officials (the voice option); secondly, they may leave districts or schools which are ineffective (the exit option). According to general incentive theories (Laffont and Martimort 2002), both options may put pressure on schools to improve. It is therefore imperative that the problems associated with the publication of these indicators, and the effects on schools of this publication, are addressed. As long as these publications continue to be issued, it is our goal to support and influence those who compute and distribute school performance indicators and, in so doing, to positively influence the impact of school performance publications on school improvement (Greene 1999, 2001; Baker *et al*. 2002).

Public Administration Vol. 88, No. 1, 2010 (90–112)
© 2010 Blackwell Publishing Ltd

This article describes a set of standards for the publication of school performance indicators that has been developed collaboratively by an international group of indicator researchers from Western Europe. These proposed standards represent an attempt to address the problems observed to be associated with the publication of school performance indicators. It is suggested that these standards may be relevant to indicator reporting in other parts of the world. We first discuss the context of the publication of school performance indicators (SPIs) and analyse the problems associated with their publication. Next, we outline the proposed publication standards and then, to illustrate their use, we apply them to some SPIs published in the US, England and The Netherlands. The article concludes with a discussion of the merits of SPIs and the value of the standards we put forward here.

THE NEW PUBLIC ACCOUNTABILITY

For many years, and in many countries, especially in Europe, neither policy-makers nor citizens demanded much accountability from teachers and schools. This situation has changed considerably during the last decade, as illustrated by three important shifts.

First, in the past, accountability transpired mainly at the political level, from which members of the general public – including those directly involved such as schools, parents and students – were, to a large extent, excluded. Accountability, in this classic sense, was a mix of political and administrative responsibility. Politicians were called to account by the electorate, and policy-makers in turn called public servants or executive bodies to account. Furthermore, politicians were primarily concerned with education at the macro level. Now, information about individual public institutions such as schools, hospitals, and other local services is available generally. As a result, voters, tax-payers and parents may look over the shoulders of politicians, public servants, and managers in the public sector, and everyone is entitled to make enquiries into public matters.

Second, early systems of school accountability focused on inputs rather than on the output of these public institutions. Schools were held accountable for ensuring that students had equitable access to high-quality instructional resources rather than being held accountable for the actual performance of the school. Much was left to the professional discretion of schools and teachers. For example, initial performance monitoring systems in the US focused on instructional resources, such as the number of textbooks available to students, average class size, and the number of fully qualified teachers. In Europe, too, school inspectors attended more to input requirements such as whether the school met the legal requirements – whether teachers had the required qualifications, for example – than to outcomes. Because the impetus for the shift towards public accountability systems originated largely from sources outside the educational establishment, from politicians and the business community, for example, new public management strategies such as performance measurement, benchmarking, and performance-linked reward schemes, became popular. The fact that technology meant that computers were available for the processing of huge datasets also played an important practical role. The nature of educational accountability changed, favouring such outcome indicators as standardized test scores and graduation rates.

Third, although accountability in the public sector has been extended legally – opportunities to demand information from the government and to take public organizations to court have increased in several countries – many systems remain hierarchical in nature. This means that individual citizens may have to take legal action to assert their rights, an

© 2010 Blackwell Publishing Ltd

type of action which many may find to be a considerable barrier. For this reason, horizontal forms of accountability, in which the public as 'client' is a significant element, are gaining in importance. 'Voice' in particular, and users' freedom of choice – 'exit', or voting with their feet – are important here (Hirschman 1970). Public reporting has become the main feature of many new accountability systems.

The need to modernize the concept of public accountability has not arisen suddenly. It is part of a general tendency to extend the market mechanism to the public sector; hence, the link between the debate about performance data publication and the various views on educational market operations (Besley and Ghatak 2003). Thus the popularity of public performance indicators may be understood when set against this background of an increase in market forces (Woods *et al*. 1998). Empirical evidence of the consequences of publishing performance indicators for schools is weak. There are relatively few studies and it is difficult to isolate the effects of publication from the effects of other reforms. During the late1980s and early 1990s in England, for example, public reporting of school performance data was an important element in the then Conservative government's efforts to create a market in the education system. The UK government at that time claimed that applying market theories and enhancing choice would lead to more effective schools by encouraging schools to perform better and to be more responsive to their customers. This claim, in particular, met with considerable criticism, resulting in a substantial body of literature on 'the market' and 'competition' in educational systems (see, for example, Gewirtz *et al*. 1995; Whitty *et al*. 1998; Woods *et al*. 1998; Hughes *et al*. 1999). To date, no convincing empirical evidence is available for the positive effects which marketization in the educational sector is assumed to have had on the effectiveness of schools.

An overview of the experiences in France and England (see Karsten *et al*. 2001) shows that parents from the middle classes, in particular, make use of the published performance data. This group of parents tends to invest more time and effort in the choice of schools for their children. Although parents consider average examination results a significant criterion for judging the quality of a school, assessment results are seldom the most important criterion for choosing a school. More important considerations include the distance to the school and the school's pedagogical climate. This type of information is typically obtained through 'word of mouth' or other informal communication among parents. The majority of the unintended side effects of SPI publication is observed within this group of parents in both England and France, especially where there are significant formal and/or material limits to educational provision. These side effects include: the use of false addresses if the choice of schools in the parents' home district is limited; the demand for homogeneous, mainstream classes; and protests against student-school allocations.

Although the proven efficacy of current publications for school improvement is subject to question, schools continue to seek strategies to improve their reputations or positions in school performance rankings. One of the best-known strategies is to regulate student intake. This strategy is more prevalent among schools which have the formal means to do so – for example, selective schools and private schools. Public schools also attempt to take this route by establishing tracking systems, or 'gifted classes'. Other strategies for boosting school performance include teaching to the test, concentrating on borderline students, and excluding 'difficult' students.

© 2010 Blackwell Publishing Ltd

PROBLEMS ASSOCIATED WITH THE USE OF SCHOOL PERFORMANCE INDICATORS (SPI) PUBLICATIONS

To investigate the subject of school performance indicators in as broad a field as possible, in 1999, the authors set up a web-based discussion about the merits and demerits of the publication of school performance indicators among a group of international experts (see Acknowledgement). As well as chairing the web-meeting, the authors raised topics for discussion. Over the course of these discussions delegates identified three broad categories of problems with regard to SPIs: (1) technical-analytical problems; (2) utilization problems; and (3) political/ethical and societal problems.

Technical-analytical problems

Technical-analytical problems comprise limitations in the validity with which SPIs reflect the quality of schools assessed. Eight are highlighted as outlined below.

1. Even if the raw achievement scores of a school's students are adjusted for relevant student characteristics as socio-economic status (SES) or prior educational attainment, precise school performance levels remain dubious if large confidence intervals are employed. It should be noted that confidence intervals are important here because the SPIs are based on a sample of students, for example, those students who took part in the central examinations for specific subjects attending the schools for which SPIs are reported. Large confidence intervals (the decision to use 95 per cent, 80 per cent, or another interval is arbitrary but it is important to note that the confidence interval gets wider as the criterion tightens) are utilized in some reporting systems, particularly when the estimation of a school's performance is based on relatively few students. Even the 10 to 20 per cent of schools for which the confidence intervals around their estimated school performance indicator do not overlap with the confidence intervals around the estimated confidence intervals of all the other 80 to 90 per cent of schools, cannot necessarily be judged poor or excellent since there may be other unidentified factors responsible for the low or high scores besides the factors included in the model. See, for example, the discussion below on student mobility.
2. In some locales, schools are assigned to 'similar schools' groups, and individual sites are compared to the group mean rather than the overall population of schools. This banding should be rejected because the basis for allocating schools to categories, for example, 'good', 'average', 'poor', is often dubious, and potential exists for schools on the boundaries of these arbitrary groups to be harmed.
3. In some instances, indicators are based mainly on data for students who entered the school several years earlier. Such indicators do not provide reliable information on the current quality of schools, and are even less insightful about future school performance.
4. Student mobility, that is, students moving to other schools or dropping out of school, causes significant difficulty in determining precisely how much value a school has added to the students it enrols. Value-added indictors reflect student progress in a given school, relative to other schools, over a certain period of time. In computing indicators, adjustments are usually made for relevant student features, for example, prior attainment scores, to allow schools to be compared fairly. A value-added school score reflects the difference between a school's score and the average school score, for

© 2010 Blackwell Publishing Ltd

schools with a similar student population in terms of average prior attainment and other relevant intake characteristics. It is quite difficult to ascertain precisely how long 'mobile' students stay in each school. If many students move in and out, and a school's final cohort is quite different from its intake cohort, robust value-added indicators are not currently possible. Taylor-Fitz-Gibbon (1997) points to similar problems with student attendance.

5. No single simple and valid measure exists for school quality (Goldstein and Thomas 1995; Ryan 2002). Schools performing equally well, based on composite measures of student achievement, may exhibit considerable internal variation from one subject to the next or one student subgroup to the next (Goldstein and Thomas 1995; Luyten and Snijders 1996). Schools may also be differentially effective for various student groups, including those based on gender and socio-economic status. For example, school effectiveness studies reveal differential effectiveness for students with varying levels of prior achievement (Scheerens and Bosker 1997). Multiple process and output indicators (Schagen and Morrison 1999) are therefore necessary for a valid school quality profile. The typical SPI system does not provide these.
6. Students with high SES backgrounds and achievement levels often enter 'better' schools. This fact may influence the effectiveness status of the schools. Some studies (see Ladd and Walsh 2002) show that value-added measures, such as those implemented in the United States in North and South Carolina, overestimate the effectiveness of schools serving higher performing students. This is not only due to 'peer group effects', but also to the failure to correct for measurement error in the test data. Other studies (see Koopman and Dronkers 1994) show that 'ceiling effects' on the test may lead to underestimations of school performance since there may be little to add to students who are already high achievers.
7. Important differences may exist between the public and private school sectors in terms of regulations, resources, recruitment of staff, and student admission standards. Comparisons between these two types of schools may therefore be unfair to public schools, unless all relevant factors are taken into account.
8. Value-added school performance indicators provide information on the relative performance of schools, but do not indicate the degree to which schools meet certain absolute educational standards (Tymms 1999).

Utilization problems

This category refers to the extent to which stakeholders, for example, schools, policy-makers, parents, and students, may utilize SPIs effectively for the purposes of school improvement, accountability, or school choice.

1. A considerable percentage of parents, particularly those from low-income backgrounds, are unfamiliar with SPIs, and have difficulty interpreting them. If SPIs are published based on the assumption that their content is valuable for the target group because the target group can benefit from the SPIs, then the wide and correct utilization of the indicators should be promoted by making them known to as many parents as possible, and by explaining their contents.
2. SPIs may not encourage staff within the highest performing schools to improve their functioning; in addition, staff within the lowest performing schools may feel discouraged, especially by non-value-added systems. In the US, some states offer

© 2010 Blackwell Publishing Ltd

recognition and/or rewards to high-growth, high-performing schools as incentives to continue improvement.
3. School staff may feel unsure about what is wrong with their schools, because general, public, SPIs do not indicate what instructional or organizational processes should be improved upon.
4. One of the most important criteria for judging a SPIs utilization is the magnitude of the benefits in relation to costs. Systems of confidential indicators, such as those implemented in England and in The Netherlands, are not as expensive as the public systems implemented throughout the US. In particular, when some of the costs of unintended effects are taken in account, public money may not be being wisely spent. Moreover, when the published performance information does not reach the majority of the intended target audience, as some European studies show (West and Pennell 2000), it is questionable whether the investment is worth the expense.

Political-ethical and societal problems

These problems consist of the unintended and unjust effects which SPIs may have on some of the actors involved, for example, unfair effects on schools and teachers, or 'window dressing' by schools.

1. Because of the potential for harm to some schools, several of our experts consider the publication of school performance data – before school effectiveness can be documented with certainty – to be unethical. Conversely, others argue that to collect school performance data and to allow some stakeholders, for example, policy-makers and schools, access to them and to deny this access to others, for example, parents, would be paternalistic.
2. In some countries, in order to promote market mechanisms in the educational sector, the publication of SPIs is accompanied by a labelling process in which some schools are identified as 'failing'. This labelling system is considered to be negative for the sector as a whole. By their very nature, inter-school comparisons always identify the schools which perform at the lowest levels in relation to the other schools as being 'at the bottom of the pile'. More relevant is the extent to which schools do achieve the absolute standards which are considered important. This argument was a leading motivation behind the rise of standards-based accountability in the US, whereby each school's performance is judged by its status in relation to a rigorous standard.
3. The publication of SPIs touches on the classic tension between evaluation and improvement. Without the external public and market pressure which is generated by public SPIs, schools may be less inclined to use performance evaluations for improvement. Examples exist, however, of confidential performance feedback resulting in improved schools (see Coe 1998; Yang *et al*. 1999). Experience in other public sector organizations shows that the publication of one or a few indicators does not automatically unleash some hidden potential for improvement within organizations. Schools are likely to vary in their response to SPIs, and such variation is strongly associated with existing school and teacher capacities. Some schools may try to ameliorate their average scores by placing greater instructional emphasis on a few areas of assessment or focus their efforts on the students who are poised to make the largest gains – the so-called 'tunnel vision' effect. In discussing the unintended consequences of publishing performance data, Smith (1995) also points to gaming behaviour, which may be defined as the deliberate manipulation of behaviour to

© 2010 Blackwell Publishing Ltd

secure strategic advantage. According to Smith, SPIs may promote such behaviour and thus divert attention from the genuine goals of schools.

DEVELOPING STANDARDS FOR SPI PUBLICATION

Although there are several problems associated with the publication of school performance data, it seems unlikely that the movement toward public accountability will end in the near future – quite the contrary. It is thus important to improve the technical quality and use of SPIs, both in terms of improving the content of the publications and minimizing their unjust and unintended effects. In our view, such improvements should include the development of publication standards. The existence of a generally accepted set of standards could help to improve public accountability systems by influencing those who publish the SPIs, and thereby positively influencing their effects. We would also hope that such standards would provide safeguards against poor quality, false or misleading labelling, misinterpretation, and other unintended negative effects.

Such standards would comprise a number of minimum requirements. If public school performance indicators fail to meet these minimum requirements, it is suggested that this raises questions about whether the SPIs can be used responsibly and whether their publication is desirable. The requirements are broad in scope, meaning that they are relevant in the many contexts in which school performance data are brought to public attention, whether on the initiative of politicians, school inspectors, governing bodies, journalists or researchers. They are considered to be general standards which apply to the publication of SPIs, whatever form they may take.

Many sources were investigated in order to ensure that the standards suggested here reflect the best current knowledge and practice. The following procedure was followed. The results of the web-meeting mentioned above (Karsten *et al*. 2001) were used to make an inventory of frequently occurring problems with, and imbalances in, the SPIs published in various countries. Based on this inventory, we identified the main clusters of themes which could be helpful in determining potential limitations of school performance indicators. For each theme, two or more experts in the field were invited to review the available information and to outline the content, feasibility and implications of SPIs for that particular theme. We used their combined contributions for a first draft; this was then discussed by a wider group of respondents at a conference held in The Netherlands of educational researchers. The participants included all Dutch tenured professors in the field of Education, the members of two Divisions of the Dutch Educational Research Association (VOR), that is, 'Educational Policy and Administration' and 'The Social Context of Schooling'. Each conference theme was first discussed in protracted conversations that concluded with discussion of the proposed standards for SPIs. In order to be as consistent as possible, the experts then revised their contributions, expanding on the insights gained at the conference, including, in cases where opinions diverged, competitive views. A comparison of the insights obtained in the conference were then compared with those from an international inventory (Visscher 2001), further supporting the endeavour to formulate standards which could be regarded as broadly acceptable. Taking the process one step further, we then weighted and summarized the insights for each theme and distilled the central criteria to be met by published SPIs. After further discussion to refine these criteria, a conference was initiated by the authors and organized by the Vereniging voor Onderwijsresearch (Organization for Educational Research), specifically to present and discuss our standards. This conference was attended

© 2010 Blackwell Publishing Ltd

by approximately 200 educational scientists and practitioners and, finally, after intensive deliberation, we obtained validation of the extent to which the standards could be regarded as minimum standards. Although we believe that our standards articulate widely accepted views on the subject, they should still be regarded as proposals. We see this article, therefore, as a further step in the deliberative process regarding the development of publication standards for SPIs. One sign of the merit/significance of our standards within the Dutch context, is the fact that the Dutch school inspectorate, having discussed our standards, proceeded to change their publications in light of them.

Our choice of categories in developing the standards was inspired by the *Program Evaluation Standards* devised by the Joint Committee on Standards for Educational Evaluation (1994), which distinguish between four categories from which we adopted three.

1. The indicators must meet strict methodological criteria. This concerns the accuracy of the indicators: their validity, reliability, and completeness.
2. Our second principle is that parents, the government and schools should benefit as much as possible from publications containing school performance indicators. For this reason the standards and recommendations for their utilization are required. Indicators should inform users and the relevant user groups should be able to understand them.
3. In our opinion, the parties involved in publishing school performance information the government, media, and researchers must not evade their responsibility with respect to the risk that this information may be used for unintended purposes and may have undesirable effects.

Thus it is recommended that various conditions of due care be met when indicators are published. These standards are meant to prevent or reduce the ethical and societal problems which were discussed above; they should be compared with the standards of propriety laid out by of the Joint Committee on Standards for Educational Evaluation (Joint Committee on Standards for Educational Evaluation 1994).

THE PROPOSED STANDARDS FOR PUBLIC SCHOOL PERFORMANCE INDICATORS

Accuracy

Value-added indicators (A1)

School performance indicators should provide insight into the 'value-added' by schools. By taking the composition of their student populations into account, performance indicators would then enable fair comparisons to be made between schools. Moreover, such indicators would reduce the selection of students who are 'profitable', in terms of social background and level of education, because the indicators reflect how much a school has added to the entrance achievement level of the students. The picture then presented would be a fairer one than that presented by the publication of raw, uncorrected average examination scores. A reliance on value-added or growth measures is consistent with the goals of NCLB.

Multilevel analysis (A2)

The indicators should ideally be based on multilevel analyses. There may be practical reasons, for example, data collection costs which impinge on this, but the most accurate

© 2010 Blackwell Publishing Ltd

results and therefore the most equitable representations of schools are obtained by means of multilevel analyses with the school level residual u_{0j} to measure the added value of schools. However, if the data do not permit such analyses, analysis of aggregate data is inevitable: an analysis including school characteristics and school averages of student characteristics. In this case, researchers should then indicate that the data do not allow the use of the best method from a scientific point of view, which in turn implies that the results must incorporate some uncertainty. Although the correspondence between a multilevel analysis and an analysis of aggregate data is relatively strong, it is flawed (see Veenstra *et al*. 1998; Bosker *et al*. 2001). If the hierarchy of data students within schools were to be taken into account, the results would be more accurate than the results of an analysis at one level only, for example, the highest level.

A more extensive discussion of the advantages of multilevel analysis over the use of aggregated data falls beyond the scope of this article (for a more detailed discussion, see Bryk and Raudenbush 1992; Snijders and Bosker 1999). A comparison of the use of aggregated data and multilevel analyses of school performance, based on data from Dutch secondary schools, showed that the method employed led to shifts for half of the schools. This illustrates that the method selected has an impact on the school performance score. Besides the accuracy of the results, another issue where the distinction between the two is particularly relevant must be highlighted: adjusting for the composition of the student population (see Veenstra *et al*. 1998; Bosker *et al*. 2001). For variables such as the abilities and cultural backgrounds of students, and the effect at student level, the within-school effect may be distinguished from the between-school effect. An analysis of school averages shows only the between-school effects. It is possible that schools with a 'difficult' student population: many children from poorly educated families, for example, compensate for this factor through a high-quality school policy so that the between-school effect of variables may be smaller than the within-school effect. As an example, assuming that the effect of the parents' education at the individual level is positive, but that schools whose students, on average, have poorly educated parents succeed in compensating perfectly for this factor. In this case, an analysis of school performance averages would indicate that the effect of the parents' education was nil. Nevertheless, a child of poorly educated parents would still, in this example, be better off in a school where the average educational level of the parents was low. In a regression analysis at school level, where schools compensate adequately for a 'difficult' student group, the schools having 'difficult' student populations are insufficiently rewarded when aggregated data are used. Adjusting for average student characteristics should be done on the basis of the within-school regression coefficients, not on the basis of the between-school coefficients.

Provision of confidence intervals (A3)

Publications that attempt to rate schools should include the confidence intervals of the indicators. This is necessary because achievement indicators are usually based on small numbers of students, which leads to a relatively high degree of uncertainty with regard to these indicators. As has been noted above, confidence intervals are important here because SPIs are based on a sample of students attending the school for which SPIs are reported. The choice of the interval width usually depends on whether the indicator is high- or low-stakes.

© 2010 Blackwell Publishing Ltd

Minimum of ten students' scores (A4)

In schools with small numbers of students, circumstances which are outside the control of the school and/or the performance of the students have a greater impact on average achievement than in schools with large student populations. This is because coincidental influences are averaged out to a lesser extent for smaller schools, that is, schools with relatively small numbers of students will exhibit accidentally high or low scores sooner than schools with many students. The number of student scores used to calculate a school performance indicator should therefore be at least ten. For each school, an average examination score is calculated based on the data of N students per school. If, for example, this average score were 7.00, we should know the standard error. If the standard deviation within the average school were 0.70, which is roughly the case in The Netherlands, the standard error would equal $0.70/\sqrt{N}$. The 80 per cent confidence interval around the school's average score would then equal 7.00 ± 1.28 standard error. If 10 students were involved, the 80 per cent confidence interval would be 6.72 to 7.28.

If large confidence intervals are employed, even if the raw achievement scores of a school's students are adjusted for relevant student characteristics, socio-economic status (SES) or prior educational attainment, precise school performance levels remain dubious. With large confidence intervals, the decision to use 95 per cent, 80 per cent, or another interval is arbitrary but it is important to note that the confidence interval gets wider as the criterion tightens. Large confidence intervals are utilized in some reporting systems, particularly when the estimation of a school's performance is based on relatively few students. Even the 10 to 20 per cent of schools where the confidence intervals around their estimated school performance indicator do not overlap with the confidence intervals around the estimated confidence intervals of all other 80 to 90 per cent of schools, cannot necessarily be judged poor or excellent. There may be other unidentified factors responsible for the low or high scores besides the factors included in the model. See, for example, the discussion above on student mobility

School rankings (A5)

Because of the uncertainties inherent in school performance indicators, school rankings which list the best performing school at the top and the worst performing school at the bottom should not in our opinion be published. Of course, we are aware that the media tend to like rankings and may not want to follow this standard, but this does not mean that they should report on rankings uncritically. Our experience in fact shows that criticizing rankings can actually help. For example, the first Dutch publication in a newspaper (*Trouw* 1997) of SPIs included school rankings; today Dutch publications give only information per school. However, to provide a grading system which would give users a rough, cursory impression of school results, policy-makers may consider placing schools in groups identified, for example, by such symbols as '++', '+', '0', '-' and '- - '. This is not a fundamental solution to the problem of uncertainty, but it makes it easier if the categories are clearly demarcated. In addition, the number of 'borderline schools' would stand out more.

Verification (A6)

It should be possible to verify the calculations used to arrive at school performance indicators. For this reason, an unambiguous description of both the data and the method of analysis should be published. It should be added that SPIs have more authority if they are transparent. However, it should be noted that it is evident that this requirement of transparency is not easy to reconcile with the use of multilevel analyses standard A2,

© 2010 Blackwell Publishing Ltd

above. The calculation procedure for multilevel-based indicators is complex. The ideal balance between accuracy and transparency depends primarily on the specific situation. Where SPIs fulfil a general accountability function, for example, as quality indicators in a benchmarking system which produces annual measures based on data from all students, the transparency requirement may be more important. However, where the implications of SPIs are crucial, for example, if they are linked to a performance-based funding system, maximally accurate SPIs are obviously desirable. Nevertheless, less than maximally accurate SPIs' correlating with true school performance can still be valuable in supporting school improvement.

Clear and careful procedures for data collection and computation (A7)
Evaluating students and school performance requires clear, careful and controllable procedures for data collection and SPI computation. In this way the kind of fraud that involves, for example, lowering the entrance scores of students purposely to increase the potential value which may be added by the school should be avoided. Again, SPIs have more authority if they are as fraud-proof as possible.

Utilization
Adequate explanation of the limitations of SPIs (U1)
Indicators should be accompanied by an adequate explanation of what information is or is not provided by their use. It should, for example, be stressed that:

- The indicators often have large confidence intervals (see standard A3, above) and that the data cannot indicate precisely how much each school has contributed to the indicator scores;
- School performance data refer to student groups and teacher teams instead of individual students and teachers, and therefore represent probabilities, not certainties;
- Quantitative data may reflect only a limited number of school quality aspects because, as yet, not all school goals may be measured quantitatively.

Clarity about the function of performance indicators (U2)
Each SPI publication should make explicit what the function of the school performance indicators is: accountability, support of school choice, and/or school improvement. Depending on the goals of the publication, appropriate performance indicators should be included.

Wide distribution (U3)
School performance indicators should be distributed to as many policy-makers, schools, and parents as possible; it is not acceptable to publish the information only on the Internet and/or in a newspaper. More effort is required to ensure that official school performance information is disseminated adequately among families with low socio-economic backgrounds and among ethnic minorities. These parent groups may be less informed about their rights, about quality differences between schools, and about 'how to work the system' than parents of high or middle class socio-economic status. The publication of SPIs, which is intended to optimize school choice processes and to strengthen the position of parents and their children, carries with it the risk of increasing existing inequalities – a so-called 'perverse effect'.

© 2010 Blackwell Publishing Ltd

User friendly indicators (U4)

In order to make them attractive for utilization by the general public, indicators should be user friendly. The additional costs which this would incur are 'earned back' by the better understanding and wider use of the SPIs. As many parents, teachers, school principals and policy-makers as possible should understand the indicators and their explanations. There is, of course, a tension between user friendliness and accuracy. The desired balance between the two depends on how accurate the SPIs need to be in a given situation (see the note on multilevel analyses in standard A2). An indicator system should provide insight into the performance of schools on each dimension certainly. But it is probable that schools differ in their scores on the various dimensions. As a result, the overall performance of a school would often be a mixture of different qualifications. For example, a school may be rated as good on indicator A, moderate on B, excellent on C, insufficient on D, and good on E and F. Such a mixture may not be comprehensible to many parents. A summary of the scores, using clear profiles which allow parents to perceive a school's strengths and weaknesses at a glance may be a solution here.

Due care

No undue preference to specific schools (D1)

Reporting on school performance should be as complete and fair as possible, and should not give undue preference to specific schools. I in order to prevent systematic errors, then, when judging students and schools, the same norms should be used for all.

Data verification rights of schools (D2)

To prevent unsystematic errors in SPI publications, schools should have the right to verify the data used for the computation of school performance indicators and the resultant conclusions before they are published.

Finding an optimum among the 13 standards referred to here is an enormous challenge. Important constraints which play a role in this are resources, legislation and political conditions; in fact, these are the same constraints encountered by the designers and publishers of the indicator systems themselves.

Dimensions of school quality

For the most part, the above standards pertain to cognitive outcomes of schooling. This reflects the emphasis, which to date has focused primarily on school results, in the domain of academic achievement. Nevertheless, we advocate that other aspects of school quality also require attention. This is important because the quality of a school – as measured by student achievement – is only one of the factors which parents take into account when choosing schools. However, it is important to discourage schools from concentrating exclusively on cognitive domain outputs. Examples of other relevant school quality dimensions include key qualifications for the rest of the student's educational and occupational career, meta-cognitive skills, social skills, and psycho-social qualities.

COMPARING INTERNATIONAL SPIS WITH THE SUGGESTED STANDARDS

In this section, school performance publications published in the USA, the UK and The Netherlands are compared with the standards outlined in the section above. This exploration is based on SPIs which were publicly available between early 2003 and mid-2004. Although these SPIs may now perform better on some standards, this is not the issue here since our choice of these SPIs is to serve as an illustration. Our goal is not so much to

© 2010 Blackwell Publishing Ltd

judge these publications, as to explore the extent to which the standards are applicable to various national contexts, and to facilitate the debate about SPI publications and their improvement.

The United States

As indicated in the introduction, most US states currently follow the practice of publishing school report cards designed to provide parents and tax-payers with information on how much students learn at a school. At first glance, the characteristics of these report cards seem to vary considerably. To illustrate the application of our standards, we used the report cards of North Carolina and Kentucky. These two states publish report cards not only to generate pressure for school improvement, but also to offer financial rewards to schools which perform well and impose sanctions on those which perform poorly. Kentucky 'schools in decline', for example, are identified through a straightforward formula calculated on the basis of quantitative growth expectations. A school will no longer be given this status once it has achieved the expected gains in test scores (David *et al*. 2000). There was no other special reason for choosing these two specific report cards, the goal simply being to illustrate the applicability of the standards.

The North Carolina School Report Card

The State of North Carolina (NC) has implemented a statewide accountability programme for elementary and secondary schools under the name 'North Carolina ABCs of Public Education'. The *North Carolina School Report Card* is part of this programme. According to the NC Department of Public Instruction, North Carolina's accountability programme has been so successful that it was used as the model for federal education legislation – the No Child Left Behind Bill introduced in 2002 by the George W. Bush administration. This legislation requires that schools show yearly progress, as measured by the percentage of students from grades 3 to 8 who are proficient in reading and mathematics. The exemplary function – the explicit link to school improvement – and the effects claimed, make the NC School Report Card suitable for a comparison with our standards (see http://www.ncpublicschools.org).

The School Report Card is produced by the NC Department of Public Instruction (NCDPI). Data collection is carried out by either the local school districts or national contractors. The NCDPI is responsible for data analysis and reporting. Quality control checks are carried out at school district, regional and state levels. The NC Report Card is important both in terms of its accountability function, and its capacity for encouraging parents to exert pressure for school improvement. Schools performing well receive financial rewards paid directly to teachers as well as distinctions which are part of a social recognition system: for example, 3.5 per cent of NC schools are designated 'school of excellence'. Poorly achieving schools are given a negative distinction: for example, 2.1 per cent are designated 'low-performing schools' and an assistance team may also be assigned to the school to help boost its performance. The ultimate censure is an intervention which may include the transfer, dismissal, or demotion of personnel.

The NC Report Card includes information about several achievement levels and a measure for 'expected growth'. The latter measure is calculated on the basis of students' prior achievement, the state average for the subject, and the grade in question, along with a correction for regression-to-the-mean effects. For example, a low score on test 1 provides more room for improvement on test 2 and vice versa. The card also differentiates among subgroups, for example by sex, race or income level.

© 2010 Blackwell Publishing Ltd

Kentucky School Report Card

The Kentucky School Report Card Project is carried out by the Kentucky Department of Education and is part of a statewide assessment programme focusing on school accountability for student achievement. Kentucky schools are required by law to publish school report cards. The information on the Kentucky School Report Card indicates 'how well our school is doing, where it is succeeding, and where there is room for improvement'. According to the web site of the Kentucky Department of Education, the purpose of the report card is 'to keep parents and community members informed about what is going on in each Kentucky school' (see http://www.kde.state.ky.us). The Report Card Statute states that the purpose of the statewide assessment programme is 'to ensure school accountability for student achievement'.

The Kentucky School Report Card is published in the form of a card for each school. It opens with a brief introduction and a chart and table representing a school's intended growth up to 2014, and the intended development of the school at two-year intervals. Besides containing quantitative data about the characteristics and achievements of the school (divided into four levels of achievement), the card provides a brief explanation of the data. The school's achievements are compared with district and state figures, as well as with national indicators. Performance indicators are not published for subgroups. A brief explanation of the profile and mission of the school is also included, as are lists of extracurricular activities, awards and improvements.

England

During the 1980s and early 1990s, the UK Conservative government substantially reformed education by introducing a more competitive, market-like environment for schools. One of the main elements of these market-oriented reforms was to provide parents, as consumers of education, with more information about the quality of schools. The so-called league tables – which rank and compare the performance of schools – were first introduced in 1992 in secondary education; in 1997, similar tables were introduced for primary schools. In 1997, the Labour government continued this policy, by emphasizing the need for parents 'to see what different schools can offer and to assess their choices realistically'.

The English school performance data (formerly called the league tables and now known as the school and college achievement and attainment tables), include results of national examinations, background data and other school characteristics. The examination results of a school are compared with those of local schools and with all other English schools. In the case of the latter, the results of 'independent schools' are also included and show how a school performs relative to other schools. Besides performance data, the English Ministry of Education (namely, the Department for Children, Schools and Families) also provides contextual data such as school type, student admission policy, sex, and age groups of students and data on other features of schools such as total student enrolment and the number of students with learning difficulties. Data on all schools in England may be found at the following web site: http://www.dfes.gov.uk/performancetables

The Netherlands

The introduction of league tables in The Netherlands was triggered by the publication of an educational supplement in *Trouw*, a national newspaper, in 1997, on the 'quality' of all secondary schools. One of the paper's journalists invoked the *Wet Openbaarheid Bestuur* (Dutch Freedom of Information Act) to request data on schools from the Inspectorate of Schools; the data were then used in the newspaper's presentation on the quality of

© 2010 Blackwell Publishing Ltd

schools. A year later, the Ministry of Education, Culture and Science published its own school performance indicator data on Dutch schools (http://www.onderwijsinspectie.nl). In secondary education, the Quality Card for secondary education has been available since 1998. This card includes the following information: name and address of the school, number of students, the type of school board ('private', that is, governed by an association or foundation, or 'public', that is, governed by the municipality), the types of secondary education offered by the school, the number of students in each type of secondary school, the average class sizes in the first and second grades, and various types of performance data, including the school's average student marks in national examinations. The cards reflect school performance relative to performance of other schools. In 2003, Quality Cards were introduced in primary education. These Quality Cards are a popularized version of the full inspection reports on primary schools. In addition, in 2003, the Quality Card for secondary education was replaced by a new version. For our illustrative pilot test, we focus on the Quality Card for secondary education, as published early in 2003.

The results of the comparison

The extent to which the publications in each of the three countries met the standards is described below, standard by standard (with the exception of 3 'factual' standards judged by the authors).

Value-added indicators

The NC School Report Card reported the percentage of proficiency and learning growth, but information was not readily available on how growth was determined. In the performance indicators in the Kentucky card, only absolute achievement levels were reported, with some pegged to standards-based achievement thresholds, and comparisons to state means and time trends.

Prior to 2001, the English school performance publications were flawed, since only raw examination results were published. In 2001, however, a pilot using value-added scores based on deviations from median expectations, predicted from average levels at the so-called previous Key Stage was successful, and such information was published nationwide in 2003. Some English experts think that the assumptions underlying the valued-added figures of the Education Department calculations and the calculations themselves are badly flawed, contain biases and random errors, and fail to take into account school circumstances. This does not mean that indicators other than value-added ones should be published since, in our view, they will be the more inaccurate.

In The Netherlands, the Inspectorate tries to provide insight into the value added by schools. The Inspectorate, however, did not have information about the performance level of students upon entry into secondary education. The secondary school recommendation depends on the aptitude of students according to test results at the end of primary school, and the opinion of their primary school teacher was therefore used as a proxy for the student's entry achievement level. This aptitude measure predicts, for example, approximately 30 per cent of the variance in examination results in the third year of secondary education. Because this proxy is used and because the value added in other indicators, for example, internal efficiency, is not estimated, the Dutch experts gave The Netherlands a low score on this standard.

Multilevel analysis

None of the indicators was based on multilevel analysis.

© 2010 Blackwell Publishing Ltd

Provision of confidence intervals
Confidence intervals were generally not available. The NC School Report Card was based on a Performance Composite for low-performing schools. A standard error was given for the composite performance index, but not for the components of the composite. The official site for the Kentucky Report Cards did not provide information on intervals or standard errors.

To find information on confidence levels on the DfES site, parents must click on a succession of explanatory notes. These gave only some examples, for example, for 50+ or 100+ students. No caveats were given for the value-added measures themselves. Most parents obtained performance data from newspapers or the BBC site, neither of which gave any information about confidence limits. In The Netherlands, confidence intervals were not included in the publications on secondary schools.

A minimum of ten students scores
The NC School Report Card indicated that a 'five or more' reporting rule was required. In the NC School Report Card, the number of students tested was provided – for the entire school and subdivided by race, economic disadvantage, and limited English proficiency – but not by grade level. For Kentucky, this information was not stated. The minimum number of student scores used to calculate a school performance indicator in both England and The Netherlands was ten.

No school rankings
There was no overall rank order in any of the publications. Merit lists were published, however. For example, North Carolina offered comparisons with the district and state at various levels of performance, and also designated 'schools of excellence', the 'ten most improved high schools', and the annual 'assignment of assistance teams'. As one of the experts noted: 'It is obvious from the data what the rank order is'. There was no ranking in Kentucky, but percentiles for the school, district, and state were reported. In England, the Education Department did not publish rankings, but the press – for example, the BBC's Education web site – did so based on the Department's figures. This was also the case in The Netherlands.

Possibility of verification
Several SPIs included information on how the calculations were made, but on most report cards there was no information about this aspect. The North Carolina card, for example, gave the address of a web site (http://www.ucreportcards.org) where more information was available. This information, however, was not sufficient for verifying the calculations, so this standard was not met unambiguously. The Kentucky card also provided some details, but in our opinion not enough to replicate the analysis. The English Department of Education described the analysis methods in technical appendices, but these were not particularly prominent. In principle, the method may be checked, but with some effort. This was also the case in The Netherlands where no direct information was available, but an interested reader could request it.

Clear and careful procedures for data collection and computation
For the two American report cards, no information was provided regarding the procedures used. The SPIs depended primarily on state test scores. As one of the experts noted:

© 2010 Blackwell Publishing Ltd

'There are no indications of fraud, nor however are there any indications of how fraud is prevented'.

Although some English experts indicated that this standard was currently not regarded as a problem, another English expert said: 'There is a widely-held belief among head teachers that, where value added is concerned, some schools depress their results at the younger age, so as to get higher value added scores at the older age, and in some cases deliberately do not appeal against external marking, giving students unexpectedly low test scores, because these low scores can help the school's overall value-added figures'. The procedures used by the Dutch Inspectorate for evaluating school performance were both clear and careful according to the experts.

Adequate explanation of the limitations of the SPIs

The NC School Report Card came with a guide which briefly explained the meaning of the performance indicators. The Internet site had extensive additional information as part of the ABCs Program: for example, about backgrounds, the test used, and the meanings of the indicators. Nevertheless, it seems questionable whether this information was sufficient for parents in terms of comprehensibility; clear explanations of how to interpret results were not offered. Although the performance indicators in the Kentucky School Report Card were accompanied by a brief explanation, and the card stated that non-quantitative measures were also important, there was no information on the limitations of the indicators for judging overall performance. Most publications of English SPIs were not accompanied by an adequate explanation of what the indicators showed and what they did not show about school performance. Only if one carefully read the notes on the DfES website might one obtain insight into these limitations. The Dutch indicators for secondary schools were accompanied by a brief explanation of what the indicators revealed about school performance.

Clarity about the function of the performance indicators

Parents were the target group of the NC School Report Cards. The NC School Report Cards referred briefly to performance targets and what constituted adequate yearly progress. However, no information was provided on how those targets were established. In addition, in the Guide to the School Report Card, the function of the indicators was not mentioned. The introductory letter, however, did refer to the accountability function. The Kentucky card was intended for parents, and included a general statement about the nature of the information, along with the remark that 'the report card is a good starting point for discussions with teachers, administrators, school council members, and school board members'. No attempt was made by the DfES to distinguish between the functions of the SPIs. The fact that printed copies of tables for secondary schools were sent to parents of students in their last year at primary school suggested that high priority was given to the school choice function. In the Dutch publication on secondary schools, it was made clear, albeit rather implicitly, that the function of the school performance indicators was to support school leaders and, above all, parents.

Wide distribution

The NC School Report Cards for each school and for all schools in the state were published on the Internet. Indicators were available for selection and comparison: for example, for a particular region and type of school. The Kentucky cards were also available on the Internet. Additionally, Kentucky schools were required by law to mail report cards to

© 2010 Blackwell Publishing Ltd

all parents, and to publish annual summaries of the report card in the largest circulation newspaper in each county. Dissemination of the card by mail in particular maximizes the probability that all parents are reached. The English indicators were published on web sites and, for secondary schools only, in newspapers. The publications of the Dutch Inspectorate were distributed mainly by means of the Internet. Parents with a child in the last year of primary education were informed of the existence of the indicators. Parents who had no access to the Internet could receive a paper copy of the information concerning the schools in their area.

User friendly indicators

The structure of the NC School Report Card was logical, but contained several terms which were difficult for many parents to understand without consulting the guide. Moreover, the card did not explain clearly how the various indicators or the weight attached to them should be interpreted. Should a parent look at year end proficiency scores? At growth? At Adequate Yearly Progress AYP targets, or at achievement alone? The North Carolina Department of Public Instruction which produced the school report card had identified the need to provide clear explanations of the information as one of the challenges for improvement. The Kentucky card was judged sufficiently user friendly for someone with good web skills. The card, however, was presented in an unattractive format: for example, closely printed text, many tables, cluttered presentation. The DfES indicators for primary and secondary schools were not attractive for use by the general public. The majority of parents, and many teachers and school governors, experienced difficulties in understanding the SPIs: 'Value-added measures are difficult to understand and most adults have poor understanding of numerical data anyway'. The Dutch secondary school performance indicators were made attractive for public use, but contained the same large amount of information.

No undue preference to specific schools

In North Carolina, schools with 'unresolved data issues' were identified and there was a list of 'schools not included'. The Kentucky and NC School Report Cards were standardized across schools, which means that the same information was provided for each school. One of the English experts said about the DfES indicators that: 'There are statistical biases in the value-added algorithms used by the DfES which penalise schools with high proportions of high-achieving students'. Hence, the indicators are unjust to this group of schools. The Dutch secondary schools were considered 'good' with respect to this standard.

Data verification rights of schools

In England, schools had the possibility to check the outcome data on which the indicators were based. In The Netherlands, the card was based on an assessment of the school by the Inspectorate. Before this assessment was confirmed, the school was given the opportunity to comment on it.

CONCLUSIONS

Having compared school performance publications from various parts of the developed world with our proposed SPI publication standards, we may now assess the current situation. Although in many cases the scores could be determined unambiguously, there

© 2010 Blackwell Publishing Ltd

were cases where our opinions varied with regard to the score given to a particular publication by our assessment panel. The ramifications of this problem are slight. In this context, it is again stressed that our concern was not with an evaluation of the report cards *per se*, but that this evaluation was used to get an impression of the utility and applicability of the proposed standards, and of the results they produced. It is hoped that the application of these standards will initiate a debate about the improvement of performance publications: in our opinion, an important step forward. In general, as assessed, the high-stakes school performance publications could be improved in many respects – and all failed to meet several of our standards.

How accurate were the published indicators? Although there seems to be a trend towards the publication of indicators which show how much value schools have added to students' entry level scores, several publications did not meet our standard of enabling fair comparisons to be made between schools by taking into account the composition of the student population (see A1, above).

The requirements to base school performance indicators on multilevel analysis (A2) and to publish confidence intervals (A3), due to the high degree of uncertainty, were not met in any of the publications assessed. As a result, it is clear that, to date, school performance has not been computed as precisely as is desirable, and the indicators imply more certainty than they can actually provide. Fortunately, most publications are based on the data of ten or more students (A4), and do not make use of school rankings (A5). Furthermore, facilities are in place to enable interested parties to verify the calculations (A6) and, in most publications, careful precautions are taken against fraud and strategic behaviour on the part of schools (A7).

Several utilization standards were not met either. The requirement that publications be accompanied by adequate explanations of exactly what information is or is not provided by the indicators (U1) was not met. Although more information was provided on the function of the school performance publications (U2), there was also considerable room for improvement in this area. In the opinion of the authors of this article, the efforts of the three countries considered in this study were, in most respects, 'reasonable'. More effort is required, however, to ensure that the publications are distributed among as many parents as possible (U3). Finally, half the publications were found not to be user-friendly (U4): they were not attractive in terms of the goal of all publications: use.

It proved quite difficult to judge the publications on due care standards. In most cases information was unavailable about the extent to which the publications gave undue preference to one or more schools (D1), or the extent to which schools were given the opportunity to verify the data before they were published (D2). The absence of this information may be due to the fact that little or no attention was given to these two standards. However, it may also be that the standards were in fact met, but that this information was simply not presented.

What can be said, finally, about the usefulness and relevance of the proposed standards? The exploration shows that the proposed methodological requirements of added value, multilevel analysis, confidence intervals, and so on were only partially met. Should this lead to the conclusion that the standards are too stringent or unrealistic, or that they should primarily be used only by scientific researchers? In our opinion, this is not the case, particularly when SPIs are used to compare schools or where the image created by SPIs has other practical consequences. Examples of these are school selections by parents and decisions by government, where careful assessment is of the utmost importance. The utilization category judgements were, typically, no higher than 'reasonable'. However,

© 2010 Blackwell Publishing Ltd

it is stressed again that the proposed standards should not be dismissed summarily as being too ambitious. SPIs may be seen as an element of a shift in the way in which education systems are managed, accounted for and improved. Transparency and market mechanisms are important incentives for supervision of the schooling system, for example, by the Inspectorate or the government and quality improvement. It is important therefore that the information on which such a system is based meets appropriate requirements: for example, clarity about what information about school quality is or is not provided by SPIs. With respect to the standards from the last category, Due Care, perhaps because of the relatively short history of school performance publications, there are few well-defined basic principles, demands and procedures which are generally accepted among educationalists, policy-makers and researchers. We suspect that the consensus about the significance of these principles and procedures will increase in due course.

The overall conclusion is that it is extremely important to invest in the improvement of both the calculation of school performance indicators and the way in which they are published. The standards may serve as a useful guide for judging and improving school performance publications. The standards have been developed to support various actors in the publication of SPIs: researchers, the media, and government. As researchers, we cannot simply say that school performance indicators should not be published because they are flawed. The fact that they are being published and will continue to be published implies that this type of educational evaluation should be further developed in order to assist in the development of improved school performance publications. Although these publications are, as yet, imperfect according to our standards, it is hoped that the publication of our standards will serve as a springboard for discussion and further improvement.

The media has an important responsibility here and should be aware that SPIs must be handled with care because they have a significant potential impact on how the performance indicators will be perceived and used. The media should therefore be urged to respect the standards. The Education Writers Association (http://www.ewa.org), based in Washington, DC, provides a good example of the role the media can play in this context (see also Benton *et al*. 2005).

As long as some standards fail to be met, governments – especially the agencies responsible for school inspections – has a responsibility to inform the target groups about the inadequacies of public reporting. Moreover, a system of accountability should be based on trust among the parties involved. If institutions 'fail', they should receive the resources – funding, expertise and information – necessary to achieve higher performance levels. It is hoped that our proposed standards for the publication of SPIs will serve as a trigger for further international debate on the final form such a set of standards should take and that we may thereby positively influence the quality and effects of school performance publications.

ACKNOWLEDGEMENT

The international experts who contributed to the web-discussion and conference on SPI standards are listed below. Members of the web-based discussion group were: R. Bosker, University of Groningen; P. Bressoux, Université de Grenoble; L. Demailly, Université de Lille; C. Fitz-Gibbon, University of Durham; H. Goldstein, University of London; S. Karsten, University of Amsterdam; D. Meuret, Université de Dijon; I. Schagen, National Foundation for Educational Research; P. Tymms, University of Durham; A. Visscher, University of Twente; A. West, London School of Economics; E. Wragg, University of Exeter.

© 2010 Blackwell Publishing Ltd

Experts invited to review the available information and comment on the themes analysed were: A. Béguin, University of Twente; R.J. Bosker, University of Groningen; J. van Bruggen, Dutch National School Inspectorate; A.B. Dijkstra, University of Groningen; S. Doolaard, University of Groningen; J. Dronkers, European University Insitute, Florence; M. van Dyck, Dutch Educational Council; W. van de Grift, School Inspectorate; P. Karstanje, University of Amsterdam; S. Karsten, University of Amsterdam; H. Leune, Educational Council; F. Mertens, School Inspectorate; H. Oosterbeek, University of Amsterdam; J. Peschar, University of Groningen; L.Rekers-Mombarg, University of Twente; E. Roede, University of Amsterdam; R. Veenstra, University of Groningen; R. van der Velden, University of Maastricht; A. Visscher, University of Twente; M. van der Wal, University of Groningen; S. Waslander, University of Groningen; D. Webbink, University of Amsterdam; B. Witziers, University of Twente; P. Zoontjens, Catholic University of Brabant. Invited to comment on the comparison between SPIs published in the United States, England and The Netherlands were: M. Berends, Vanderbilt University; J. Dronkers, European Institute in Florence; A. Gamoran, University of Wisconsin, Madison; W. Meijnen, University of Amsterdam; H. Oosterbeek, University of Amsterdam; D. Webbink, CPB Netherlands Bureau for Economic Policy Analysis, The Hague; J. Critchlow, Principal, Bedales High School, North Yorkshire; C. Dean, University of Durham; I. Schagen, National Foundation for Educational Research; I. Selwood, University of Birmingham. The authors alone take full responsibility for the views contained in this article.

REFERENCES

Apple, M. 2001. 'The Rhetoric and Reality of Standards-Based School Reform', *Educational Policy*, 15, 601–10.

Baker, E.L., R.L. Linn, J.L. Herman and D. Koretz. 2002. *Standards for Educational Accountability Systems*, Police Brief No. 5. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Benton, J., H. Hacker and H. Booth. 2005. 'State plan TAKS cheating inquiry', *The Morning Dallas News*, 1 October 2005.

Besley, T. and M. Ghatak. 2003. 'Incentives, Choice, and Accountability in the Provision of Public Services', *Oxford Review of Economic Policy*, 19, 2, 235–49.

Bosker, R., A. Béguin and L. Rekers-Mombarg. 2001. 'Hoe meten we de prestatie van een school?' [How do we measure school performance?], in A.B. Dijkstra, S. Karsten, R. Veenstra and A. Visscher (eds), *Het oog der natie: scholen op rapport; Standaarden voor de publicatie van schoolprestaties*. Assen: Van Gorcum, pp. 121–35.

Bryk, A.S. and S.W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Coe, R. 1998. *Feedback, Value-added and Teachers' Attitudes: Models, Theories and Experiments*, doctoral thesis. Durham: University of Durham.

David, J., P. Kannapel and C. McDiarmid. 2000. 'The Influence of Distinguished Educators on School Improvement', *A Study of Kentucky's School Intervention Program*. Lexington, KY: Partnership for Kentucky Schools.

Foxman, D. 1997. *Educational League Tables: For Promotion or Relegation*. London: ATL Publications.

Gewirtz, S., S. Ball and R. Bowe. 1995. *Markets, Choice and Equity in Education*. Buckingham: Open University Press.

Goldstein, H. and S. Thomas. 1995. 'School Effectiveness and ''Value-added'' Analysis', *Forum*, 37, 2, 36–8.

Greene, J.C. 1999. 'The Inequality of Performance Measurements', *Evaluation*, 5, 160–72.

Greene, J.C. 2001. *Beyond Accountability*. Keynote address at the annual meeting of the Southeast Evaluation Association, Tallahassee, FL.

Hirschman, A.O. 1970. *Exit, Voice, and Loyalty*. Cambridge, MA: Harvard University Press.

Hughes, D., H. Lauder, H. S. Watson, *et al*. 1999. *Trading in Futures: Why Markets in Education Don't Work*. Buckingham: Open University Press.

Joint Committee on Standards for Educational Evaluation. 1994. *The Program Evaluation Standards*. Thousand Oaks, CA: Sage.

Karsten, S., A. Visscher and T. de Jong. 2001. 'Another Side to the Coin: The Unintended Effects of the Publication of School Performance Data in England and France', *Comparative Education*, 37, 2, 231–42.

Koopman, P. and J. Dronkers. 1994. 'De effectiviteit van algemeen bijzondere scholen in het algemeen voortgezet onderwijs' [The effectiveness of general private secondary schools], *Pedagogische Studiën*, 71, 420–41.

Ladd, H.F. and R.P. Walsh. 2002. 'Implementing Value-Added Measures of School Effectiveness: Getting the Incentives Right', *Economics of Education Review*, 21, 1, 1–17.

© 2010 Blackwell Publishing Ltd

Laffont, J.J. and D. Martimort. 2002. *The Theory of Incentives: The Principal-Agent Model*. Princeton, NJ: Princeton University Press.
Luyten, H. and T. Snijders. 1996. 'School Effects and Teacher Effects in Dutch Elementary Education', *Educational Research and Evaluation*, 2–1, 1–24.
Peterson, P.E. and M.R. West (eds). 2003. *No Child Left Behind? The Politics and Practice of School Accountability*. Washington, DC: Brookings Institution.
Ranson, S. 2003. 'Public Accountability in the Age of Neo-liberal Governance', *Journal of Education Policy*, 18, 5, 459–80.
Ryan, K. 2002. 'Shaping Education Accountability Systems', *American Journal of Evaluation*, 23, 4, 453–68.
Schagen, I. and J. Morrison. 1999. 'A Methodology of Judging Departmental Performance within Schools', *Educational Research*, 41, 1, 3–10.
Scheerens, J. and R.J. Bosker. 1997. *The Foundations of Educational Effectiveness*. Oxford: Elsevier Science.
Smith, P. 1995. 'On the Unintended Consequences of Publishing Performance Data in the Public Sector', *International Journal of Public Administration*, 18, 277–310.
Snijders, T.A.B. and R.J. Bosker. 1999. *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modeling*. Newbury Park, CA: Sage.
Taylor-Fitz-Gibbon, C. 1997. *Final Report of the Value-added National Project*. London: School Curriculum and Assessment Authority.
Trouw (newspaper) 25 October 1997. *'School Performance'*. Amsterdam: PGM.
Tymms, P. 1999. *Baseline Assessment and Monitoring in Primary Schools*. London: Fulton Publishers.
Veenstra, R., A.B. Dijkstra, J.L. Peschar and T.A.B. Snijders. 1998. 'Scholen op rapport. Een reactie op het *Trouw*-onderzoek naar schoolprestaties [Schools on report. A comment to the *Trouw* study into school performance], *Pedagogische Studiën,* 75, 121–34.
Visscher, A.J. 2001. 'Public School Performance Indicators: Problems and Recommendations', *Studies in Educational Evaluation*, 27, 199–214.
West, A. and H. Pennell. 2000. 'Publishing School Examination Results in England: Incentives and Consequences', *Educational Studies*, 26, 4, 423–36.
Whitty, G., S. Power and D. Halpin. 1998. *Devolution and Choice in Education: The School, the State and the Market*. Buckingham: Open University Press.
Woods, P., C. Bagley and R. Glatter. 1998. *School Choice and Competition: Markets in the Public Interest?* London: Routledge.
Yang, M., H. Goldstein, T. Rath and N. Hill. 1999. 'The Use of Assessment Data for School Improvement Purposes', *Oxford Review of Education*, 25, 469–83.

Date received 18 February 2008. Date accepted 8 June 2008.

© 2010 Blackwell Publishing Ltd