SAVE OUR SCHOOLS

Research Paper

"Naively Stomping Around in an Uninformed Epistemopathological Fog"

The Case for Reporting Statistical Error on School Results

Trevor Cobbold

March 2009

http://www.soscanberra.com

Summary[1]

According to the new Melbourne Declaration of national goals for schooling, Australian governments are committed to ensuring that information published on school results will be accurate and fair. However, to date, governments have not indicated what action will be taken to implement this commitment. The current draft Action Plan on the Declaration merely repeats the commitment. It needs to do better than this to ensure accuracy and fairness in reporting school results.

Considerable uncertainty surrounds the accuracy and reliability of school results because of measurement and sampling error. These errors are inevitable in testing and reporting regimes.

Many technical studies of school results and school league tables demonstrate that chance differences account for a significant proportion of the differences in school test scores. In the case of gains from one year level to the next or annual changes in the results of a given year level, the margin of error can be exceptionally large. Several studies, including one Australian study, show that the results of up to 80% or more of schools are indistinguishable from the average school outcome. Real differences in school results can be only identified for a small minority of schools.

This level of error wreaks havoc when comparing school results. It is not possible to make reliable comparisons or rankings of schools because they may reflect chance differences in school performance rather than real differences. Such comparisons are mostly identifying lucky and unlucky schools, not good and bad schools.

To date, this has not been recognised in government statements on reporting school results. Yet, it has major implications for government objectives in reporting school results and could have pernicious consequences for many schools.

Parents may be misled in choosing a school. Some schools may be recognised as outstanding while others are identified as unsuccessful simply as the result of chance and not because of actual programs and teaching practice. It also means that current school performance is highly misleading as a guide to future school performance.

Choosing schools on the basis of a chance variation in a school's results can have far reaching consequences. A school may be wrongly labelled in the public eye as unsuccessful and exiting parents may initiate a spiral of decline in an otherwise successful school.

Random errors in school results also make it difficult to identify effective school practices. It may mislead decision-makers and schools in recommending and adopting particular educational programs. Action taken to assist less successful schools may appear more effective than it is in practice.

The large degree of uncertainty about school results also makes the Prime Minister's threat to reward or sanction schools based on their published test results especially fraught. It is likely to result in much unfairness in the treatment of schools and their staff. In particular, using test

[1] The title of this paper is due to Rowe 2004: 13. See p.8.

results to reward teachers on the basis of student progress creates considerable potential for rewards to be misdirected and not actually reward good teaching.

Reporting the results of individual schools should be statistically valid and reliable so as not to mislead parents, schools and the public. Australian governments should commit to reporting margins of error for each school result including the associated score range. Published tables of school results should contain ‘health warnings’ about the limitations of the data and advice on how to interpret the data.

These steps to better inform the public about the accuracy and reliability of school results have been recommended by expert statistical authorities such as the Statistics Commission and the Royal Statistical Society in the United Kingdom as well as the National Center for Research on Evaluation, Standards and Student Testing in the United States.

Introduction

The Melbourne Declaration on national goals for schooling released at the end of last year commits governments to reporting school results. It further states that "in providing information on schooling, governments will ensure that school-based information is published responsibly, so that public comparisons of schools will be fair...." and "accurate".

The draft Action Plan to implement this and other commitments in the Declaration states that governments will take action to "implement fair, public, comparable national reporting on individual school performance..." In other words, the draft Action Plan failed to come up with anything to ensure responsible and accurate reporting of school results.

In one sense, the drafters of the Action Plan faced an impossible task. The research evidence from overseas on league tables is that they are never fair and that they promote social segregation and inequity in education [Cobbold 2009]. For these reasons, league tables of school results should be rejected.

However, it is clear that governments are pressing ahead with the introduction of reporting the results of individual schools. This means that the publication of league tables is inevitable. The data should at least be reported accurately and not in a way that misleads parents and the public about school performance. Education ministers need to come up with something more than simply repeating the glib assurances of the Melbourne Declaration about fairness and accuracy in reporting.

A fundamental commitment should be to report statistical errors on school results. Reporting statistical error explicitly displays the extent of uncertainty associated with school results and rankings. Failure to report statistical error for school results leads to misleading comparisons of school performance because apparent differences between schools may be over-stated or insignificant.

Failure to take account of statistical error in reporting school results raises fundamental issues about how official data on school results is used. As one of the pre-eminent scholars in education statistics stated recently:

> The league table culture is symptomatic of a deeper problem with public debate that should concern citizens. Namely, a surface precision associated with numerical data is used, sometimes unscrupulously, sometimes in ignorance, as a substitute for serious and well-informed debate. The promotion of school league tables as if they convey uncontestable information about schools is just one example... [Goldstein 2008: 398]

This paper explains the sources of uncertainty around reported school results and reviews a wide range of studies of the extent of this uncertainty. It examines the implications of uncertainty about school results for parent choice of schools, public judgements about the success or otherwise of schools and for education policy decisions. It recommends that Australian governments should commit to reporting statistical errors on published school results and warn the public about the proper use of this data.

Measurement and sampling error in test results

All tests incur random errors or non-persistent chance variation in student results which cause differences in school results that do not reflect differences in actual performance. These errors consist of measurement and sampling errors.

Measurement error is a result of inconsistency in test results because the same students may achieve different results on the same test on different days because of differences in their own well-being, such as lack of sleep or food, or because of variations in external factors such as how cold or hot conditions are in the room in which the tests are conducted. It also arises from differences in the items selected for testing and the way answers are scored.

Sampling error arises from differences in the selection of students to participate in tests. A group of students selected for a test are likely to achieve different results from another group simply because of differences in their composition. The group selected for testing may not reflect the average level of ability of all students. The smaller the sample, the more likely there will be a significant difference between the average results of the sample tested and the results if all students were tested.

Sampling error occurs even when all students in a year cohort are tested. This is because inferences are made about school performance by testing selected cohorts, such as Years 3, 5, 7 and 9 in the national literacy and numeracy assessments. Each cohort of students tested is a sample of the students in the school for the purpose of measuring school performance. As one testing expert explains:

> This question was a matter of debate among members of the profession only a few years ago, but it is now generally agreed that sampling error is indeed a problem even if every student is tested. The reason is the nature of the inference based on scores. If the inference pertaining to each school.....were about the particular students in that school at that time, sampling error would not be an issue, because almost all of them were tested. That is, sampling would not be a concern if people were using scores to reach conclusions such as 'the fourth-graders who happened to be in this school in 2000 scored higher than the particular group of students who happened to be enrolled in 1999.' In practice, however, users of scores rarely care about this. Rather, they are interested in conclusions about the performance of schools. For the inferences, each successive cohort of students enrolling in the school is just another small sample of the students who might possibly enroll, just as the people interviewed for one poll are a small sample of those who might have been. [Koretz 2008: 170]

While there is still debate on this issue, the important point is that measurement error is present regardless of whether students are treated as a sample or as a population and that in some circumstances measurement error can be as substantial as sampling error [Betebenner et.al. 2008].

The extent of error in testing varies according to the sample size. As the sample size increases the standard error decreases. Small samples result in larger errors. This is particularly relevant to reporting school results where the years tested include very few students. For example, many schools in Australia have only 25-30 students or less in the years tested under the national assessment and reporting program.

The important point is that testing students and reporting school results inevitably involves a certain amount of inaccuracy.

Measurement inaccuracy is an inevitable characteristic of measurement. If we accept the need for educational measurement then we must accept the inevitability of error. There is no such thing as perfection when it comes to validity, reliability or comparability. [Newton 2005: 436]

It is also important to understand the potential extent of this inaccuracy and uncertainty about school results as well as its implications for decisions about schooling.

Chance accounts for much of the differences in school test results

The presence of statistical error means that a school score on a test for a Year level cannot always be interpreted to represent the actual level of performance. Many studies of school performance reporting in England, the US and Australia have shown that a large proportion of school results are statistically indistinguishable when measurement error is taken into account. This has been demonstrated when raw scores are used to compare schools, when estimates are made of the school contribution to student progress (the so-called 'value added' by a school) and when estimates of value added take account of student background factors such as gender, socio-economic status, ethnicity, special education needs and students with English as a second language (so-called 'contextual value added').

Key US studies show that random factors accounted for almost 15% of the total variation in average fourth-grade test scores for combined reading and maths across North Carolina elementary schools, almost 50% of the total variation between schools in gains in scores during fourth grade, and a massive 73% of the variation in annual changes in fourth-grade scores [Kane & Staiger 2002a; see also Kane & Staiger 2002b]. These results were for schools with an average of 56 students per Year level.

A study of the proportions of students performing at different achievement levels in grades 4-8 in over 400 school districts in Iowa found that sampling error accounted for about two-thirds of the observed variability of estimates of change in proportions from one year to the next [Arce-Ferrer et.al. 2002]. Other sources of error, such as measurement error and equating error, and intervention effects jointly accounted for about one-third of the observed variability in estimates of change in proportions.

These and other research findings demonstrate that comparing results from one year to the next is highly unreliable because of chance factors [see also Hill & DePascale 2003; Colardarci 2003; Zvoch & Stevens 2006].

As expected, the extent of random errors varies according to school size. Kane & Staiger [2002a] found that over various school sizes (for Year levels tested) about 50-80% of changes in average school scores from year-to-year in North Carolina schools was due to random or non-persistent factors.

Sampling error is a significant issue in comparisons of school results which include small schools. The sampling error in the results of small schools can be quite large and leads to unreliable comparisons of school results. Average scores are more unreliable because the number of tested students is small. The results can be heavily influenced by those of 4 or 5 students in each year level tested. Large schools are less affected than small schools by differences in the student body from year-to-year.

Kane & Staiger [2002a; see also Kane et.al. 2002] found that for the smallest one-fifth of North Carolina elementary schools (average of 28 students per Year level), random factors

accounted for 20% of the total variation in average fourth-grade test scores, 58% of the total variation in gains in scores during fourth grade, and 79% of the variation in annual changes in fourth-grade scores. For California schools, 86% of the year-to-year changes in test scores for small schools were due to random causes.

One effect of this is that small schools can be disproportionately represented in the groups of schools that make the most improvement and the least improvement from year-to-year [Linn & Haug 2002: 35].

Margins of error

One way to gauge the significance of statistical error is to report it as a margin of error, or degree of uncertainty associated with the result. The margin of error is estimated as the range of scores in which there is a reasonably high probability that the true score lies.

Several technical studies have estimated the extent to which school results are different from the overall average (whether for the mean score or the average gain over time) and the size of the uncertainty interval for each school. Intervals that overlap the overall average score indicate no statistical difference in school performance.

Using data on average test scores from 48 English junior schools, Goldstein [1997] found that that the uncertainty intervals indicated that the results of three-quarters of the schools were not statistically different from the average. A further study of average test scores for 76 primary schools in Hampshire found that the uncertainty intervals overlapped for 80 per cent of schools [Goldstein et.al. 1999]. A recent study has shown that in one selected local education authority in England the average scores of 50% of secondary schools were statistically indistinguishable from the overall average for all schools in the authority [Goldstein & Leckie 2008; see also Benton et.al. 2003].

Studies of school performance using value added estimates also indicate that the performance of a majority of schools is statistically indistinguishable. One study using data on 75 English secondary schools found that the uncertainty intervals for value added estimates overlapped for three-quarters of the schools [Goldstein & Thomas 1996; see also Goldstein 1996]. A report sponsored by the UK Department of Education on value added results for 16-18 year-old students in 364 schools found that progress for the highest achieving students was not statistically different from the average in 75% of the schools [O'Donoghue et.al. 1997]. The same study found that progress for the highest and lowest achieving students was not statistically distinguishable from the average in 66% of the schools.

Another recent study of contextual value added by nearly 3000 English schools concluded that school rankings that take account of student background factors are largely meaningless as the value added estimates of almost half of English secondary schools are indistinguishable from the national average [Wilson & Piebalga 2008].

Similar results have been obtained from US studies. A study of fourth grade reading and math results of elementary schools in North Carolina found that among schools near the national average in size (between 65 and 75 students with valid test scores), the margin of error (uncertainty interval) extended from approximately the 25th to the 75th percentile [Kane & Staiger 2002a]. That is, it wasn't possible to distinguish between the average results of 50% of schools.

A study of school value added results in 60 California schools concluded that it was only possible to reliably distinguish the top third of schools from the bottom third [Betebenner 2004]. This means it was not possible to distinguish the results of the bottom third of schools from the middle third, or those of the top third from the middle third. Another study has showed that the uncertainty intervals for the proportion of students achieving proficiency levels in small schools can be very wide in comparison with large schools [Coladarci 2003].

In Australia, a study conducted by the Australian Council for Educational Research found that 84 per cent of the uncertainty intervals for school reading scores in PISA 2000, adjusted for SES and gender, overlapped the average score of all schools [Rowe 2004]. As the author of this study states:

....it illustrates that attempts to separate or rank schools in the form of 'league tables' are subject to considerable uncertainty....Interpretation of estimates of individual schools is problematic, misleading and potentially irresponsible. Unfortunately, similar to their counterparts in the UK and the USA, Australian politicians and senior bureaucrats currently advocating the publication of such PI [performance indicators] 'league tables', *are naively 'stomping around' in an uninformed epistemopathological fog.* [13]

The uncertainty problems associated with comparisons of school results was emphasised in a recent study of the extent to which current school performance can be used as a guide to future performance, using the results of 266 secondary schools across England [Leckie & Goldstein 2009]. The study estimated school outcomes at the end of secondary school for the current intake cohort. Predicting future performance on the basis of current performance adds another layer of uncertainty into school comparisons. It found that the uncertainty intervals for the predicted value added estimates were 3.5 times as wide as those for the current value added estimates. As a result, it found that almost no schools are significantly different from the average school and very few schools can be predicted to be significantly different from each other. These results are likely to be even stronger for primary schools since these are, on average, much smaller than the average secondary school.

Implications for school comparisons

The main arguments in support of reporting school results are that it will help parents choose the right school for their children; help determine successful education practices; and help in deciding whether to take action on schools not succeeding. Information used for these purposes has to be quality information. Above all, it must be reliable and capable of making accurate distinctions between the results of different schools.

The evidence from overseas studies of reporting school results and league tables is that reporting average school results, gains from one year level to another and changes in the results of year cohorts from year-to-year are unreliable and inaccurate. Study after study has shown that there is so much statistical uncertainty attached to school performance results that differences in school results can be only identified for a minority of schools. In England, up to two-thirds of the average results of primary schools cannot be separated and in both England and the United States changes in the results of given year levels from one year to the next are not distinguishable in up to 80 per cent of schools.

… we now know from a number of studies that estimates of these contributions have so much statistical uncertainty attaching to them that it is impossible reliably to make valid comparisons between most schools. It is this finding above all that provides strong evidence against the current policy on the publication of league tables – of whatever kind, and those who advocate evidenced based policy making

would do well to understand this….schools can only be separated statistically if their intervals do not overlap [Goldstein 1998].

In addition, standard errors and hence confidence intervals for these residuals were estimated. These demonstrate a great deal of overlap between schools, and call into question the idea of 'rank-ordering' schools on the basis of this kind of analysis. Schools with particularly high or low residuals may be identified, but the majority could be arranged in an arbitrary order without conflicting with the data. [Benton et.al. 2003: 75]

...the confidence intervals (how big the range of possible league-table positions for any school must be before we are 95 per cent sure that the correct one is in there) are still so large that we cannot really tell most of the schools apart, even though they will move around with much drama from one year to the next in the published tables. [Blastland & Dilnot 2008: 188]

Thus, random errors in school test scores play havoc with comparisons of school results and ranking systems. It means that it is not possible to make reliable comparisons or ranking of schools because they may reflect chance differences in school performance rather than real differences. The official publication of school test results "...lends any comparisons based upon them an authority they do not possess" [Goldstein & Spiegelhalter 1996: 405].

In particular, random errors in school results mean that school performance and school rankings are highly unstable from year-to-year. It is highly misleading to compare changes in school performance from one year to the next, especially in the case of smaller schools. It leads to unwarranted conclusions about changes and often unfairness in the inferences drawn about schools. Such comparisons "are mostly identifying lucky and unlucky schools, not good and bad schools" [Grissmer 2002: 269].

One implication is that government objectives in reporting school results outlined above are likely to be not met. Random errors in test results and the resulting volatility in test score measures will send confusing and misleading signals to parents and schools about which schools are successful and which educational strategies are worth pursuing [Kane & Staiger 2002a: 253].

Indeed, the consequences could be quite pernicious. Parents could be led into making choices that adversely affect their children. Education practices and programs could be falsely identified as successes while successful programs in reality are ignored or even falsely condemned. Schools and teachers could be rewarded or punished for chance results.

The Prime Minister has stated that the system of reporting school results is designed to encourage parents to vote with their feet [Rudd 2008b]. The Federal Education Minister says that it will better inform parent choice of school. However, a failure to report statistical error and explain the extent and meaning of uncertainty in reported school results could mislead many parents in choosing a school and in deciding whether to shift to another school. Some schools may be recognised as outstanding while others as not successful simply as the result of chance and not because of actual practices. In particular, using current school performance as a guide to future school performance is highly misleading.

One justification for the publication of school league tables is that they are able to inform parental school choice. However, these tables do not adjust for prediction uncertainty, nor do they provide a clear statement of this statistical shortcoming. Importantly, when we account for prediction uncertainty, the comparison of schools becomes so imprecise that, at best, only a handful of schools can be significantly separated from the national average, or separated from any other school. This implies that publishing school league tables to inform parental

school choice is a somewhat meaningless exercise. [Leckie & Goldstein 2009: 16; see also Goldstein & Leckie 2008: 69]

It also means that a chance variation in a school's results can have far reaching consequences, particularly in the case of smaller schools. For example, schools may be condemned as failing in the public eye as a result of a large chance decline in average results while others may be incorrectly lauded as achieving outstanding results because of a chance result.

Moreover, a large chance decline in a school's results in one year and the consequent slump in performance compared to other schools may cause parents to shift their children to another school. If, as the research indicates, the families making these choices are from a higher socio-economic status background their exit could lead to a real decline in the school's test scores because students of these families have, on average, higher levels of achievement. Thus, a chance result could initiate a spiral of decline for some schools.

A further implication of not reporting statistical errors on school results is that decisions about the success of different education programs may also be based on misleading information. The Federal Education Minister has placed considerable weight on comparisons of so-called like schools to identify successful school practices. She has argued that school results for like-schools could be used to identify best practice and innovation in successful schools [Gillard 2008].

However, the Minister's faith in the ability to identify successful schools by their school test results is misplaced. Failure to take account of statistical error may lead to some school programs and practices being falsely judged as successful while programs that are successful are not identified as such. Schools may be misled into changing otherwise successful programs and teaching methods because of a chance result while others retain less successful practices under the false impression that they have been a success. This potential consequence has been emphasised by several studies. For example:

Standard errors of the year-to-year change in proportions are sufficiently large that, in situations involving single tests and single year-to-year comparisons, practitioners are likely to be making decisions about the efficacy of educational programs based on random sampling effects rather than actual school intervention effects. This is particularly true for estimates of change based on single-year proportions and small grade-group sizes. [Arce-Ferrer et.al. 2002: 70]

It is likely to be a mistake to assume that the practices of the schools recognized as outstanding are ones that should be adopted by other schools.....
It also means that strategies of looking to schools that show large gains for clues of what other schools should do to improve student achievement will have little chance of identifying those practices that are most effective. On the other hand, schools that are identified as "in need of improvement" will generally show increases in scores the year after they are identified simply because of the noise in the estimates of improvement and not because of the effectiveness of the special assistance provided to the schools or pressure that is put on them to improve. [Linn & Haug 2002: 34, 35]

....to the extent such rankings are used to identify best practice in education, virtually every educational philosophy is likely to be endorsed eventually, simply adding to the confusion over the merits of different strategies of school reform. [Kane & Staiger 2002a: 236]

The large degree of uncertainty about school results also makes the use of rewards and sanctions based on performance especially fraught.

Such volatility can wreak havoc when rewards and punishments are doled out on the basis of changes in test scores; school personnel are at risk of being punished or rewarded for results that are beyond their control. [Kane et.al. 2002: 60]

For example, it is commonly suggested that teachers should be rewarded on the basis of student progress at school. Given the large uncertainty associated with measuring student progress in a school and a class, there is considerable potential for rewards to be misdirected and not actually reward good teaching. Similarly, the Prime Minister's suggestion [Rudd 2008a] that principals and senior staff could be replaced in schools that are not improving their performance is likely to result in much unfairness in the treatment of schools and their staff if uncertainty about the results is not taken into account.

Kane & Stager [2002a] show that rewards for schools achieving very high scores and sanctions against those achieving very low scores primarily affect small schools and have very little impact on large schools. Small schools tend to be over-represented amongst those with extremely high and low test scores because of the large variance in the results of small schools.

School reporting should be statistically valid and reliable

As the Melbourne Declaration states, it is crucial that information on school performance be used responsibly. Reports on school performance should be statistically valid and reliable so as not to mislead parents and the public.

Given that schools will be judged on their published results it is imperative that random errors affecting these results should be reported and taken into account in drawing inferences about schools, their programs and teachers. Reporting of school results should be accompanied by information about the dependability of those results. The "margin of error" or uncertainty intervals should be reported for all school scores in order to provide a more accurate indication of what the true score is and so as not to mislead about school performance. The presentation of uncertainty intervals should be as prominent as that of the values of the performance measures themselves. One way of doing this is to present the range of scores, or the range of changes in scores, indicated by the uncertainty intervals.

Many key statistical studies of published school results and league tables recommend that measures of school performance should be accompanied by estimates of statistical uncertainty [for example, Goldstein & Myers 1996]. Several overseas expert statistical authorities have also recommended that statistical errors on school results be reported as a range of scores. For example, the Statistics Commission of the UK has supported the publication of school scores as a range of scores that reflect the extent of uncertainty [Statistics Commission 2004: para 19].

The Royal Statistical Society of the United Kingdom has recommended that performance measurement data should always include measures of uncertainty and that the uncertainty of rankings should be indicated by the use of plausible ranges of rank.

League tables without demonstration of ranking uncertainty should be avoided and, even with uncertainty incorporated, such tables should be used with considerable caution. [Bird et.al. 2005: 16]

The National Center for Research on Evaluation, Standards and Student Testing in the United States has also recommended that the margin of error on school results should be reported.

Reports of results both for individual students and for schools should be accompanied by information about the margin of error in the results. Reporting probabilities that a student or school is misclassified as the consequence of the assessment's measurement error is a good way of conveying the degree of uncertainty that is associated with assessment results. [Linn 2001a: 31; see also Linn 2001b: 4, Baker & Linn 2002: 21, 24]

It is also imperative that parents, school staff, education decision-makers and the public understand the inherent limitations of school test results in order to be able to draw proper inferences.

....it is crucial for those who use test and examination results to understand what kind of inferences can legitimately be drawn from them and what kind of inferences cannot. [Newton 2005: 431]

Parents have a right to information about potentially misleading inferences which can be drawn from school results and league tables. The publication of school results should be accompanied by advice about the limitations of the data and how to interpret the data.

A 'health warning' seems to us essential and we believe that government has a major responsibility for ensuring that this is done. [Plewis & Goldstein 1998]

In summary, Australian governments should commit to reporting margins of error for each school result including the associated score ranges, together with a clear explanation of the implications of uncertainty about the results for judgements about the quality of schools and changes in performance. A 'health warning' should be prominently displayed on all published tables of school results.

References

Arce-Ferrer, Alvaro; Frisbie, David A. and Kolen, Michael J. 2002. Standard Errors of Proportions Used in Reporting Changes in School Performance With Achievement Levels. *Educational Assessment*, 8 (1): 59-75.

Baker, Eva L. and Linn, Robert L. 2002. Validity Issues for Accountability Systems, CSE Technical Report 585, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, December. Available at: http://www.cse.ucla.edu/products/rsearch.asp

Benton, Tom; Hutchison, Dougal; Schagen, Ian and Scott, Emma 2003. Study of the Performance of Maintained Secondary Schools in England. Report for the National Audit Office, National Foundation for Educational Research, November. Available at: http://www.nfer.ac.uk/publications/other-publications/downloadable-reports/study-of-the-performance-of-maintained-schools-in-england.cfm

Betebenner, Damian 2004. An Analysis of School District Data Using Value-added Methodology. Report 622, Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, University of California, Los Angeles, March. Available at: http://www.cse.ucla.edu/products/summary.asp?report=622

Betebenner, Damian; Yi Shang, Yun Xiang, Yan Zhao and Xiaohui Yue 2008. The Impact of Performance Level Misclassification on the Accuracy and Precision of Percent at Performance Level Measures. *Journal of Educational Measurement*, 45 (2): 119-137.

Bird, Shiela M.; Cox, David; Farewell, Vern T.; Goldstein, Harvey; Holt, Tim and Smith, Peter C. 2005. Performance Indicators: Good, Bad and Ugly. *Journal of the Royal Statistical Society* A, 168 (1): 1-27.

Blastland, Michael and Dilnot, Andrew 2008. *The Tiger That Isn't: Seeing Through a World of Numbers*. Profile Books, London.

Cobbold, Trevor 2009. League Tables. *The Professional Educator*, March.

Coladarci, Theodore 2003. Gallup Goes to School: The Importance of Confidence Intervals for Evaluating "Adequate Yearly Progress" in Small Schools. Rural School and Community Trust, Washington DC, October. Available at: http://www.ruraledu.org/site/apps/nlnet/content3.aspx?c=beJMIZOCIrH&b=2820553&content_id=%7BA0A3F1D1-02C4-4F39-A0D6-2D9678E56543%7D&notoc=1

Gillard, Julia 2008. Speech to ACER Research Conference, Brisbane, 11 August.

Goldstein, Harvey 1996. Relegate the Leagues: Data From Performance Tables is Crude and Often Misleading. *New Economy*, 3 (4): 199-203.

Goldstein, Harvey 1997. Methods in School Effectiveness Research. *School Effectiveness and School Improvement*, 8 (4): 369-395.

Goldstein, Harvey 1998. Models for Reality: New Approaches to the Understanding of Educational Processes. Lecture at the Institute of Education, University of London. Available at: http://www.cmm.bristol.ac.uk/team/HG_Personal/models%20for%20reality.pdf

Goldstein, Harvey 2008. Evidence and Education Policy - Some Reflections and Allegations. *Cambridge Journal of Education*, 38 (3): 393-400.

Goldstein, Harvey and Leckie, George 2008. School League Tables: What Can They Really Tell Us? *Significance*, June, 67-69. Available at: http://www.cmm.bristol.ac.uk/team/HG_Personal/Full%20Publications%20-%20download/Table%20of%20publications.htm

Goldstein, Harvey and Myers, Kate. 1996. Freedom of Information: Towards a Code of Ethics for Performance Indicators. *Research Intelligence*, July, 12-16.

Goldstein, Harvey and Spiegelhalter 1996. League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance. *Journal of the Royal Statistical Society* Series A, 159 (3): 385-443.

Goldstein, Harvey and Thomas, Sally 1996. Using Examination Results as Indicators of School and College Performance. *Journal of the Royal Statistical Society* Series A, 159 (1): 149-163.

Goldstein, Harvey; Huiqi, Pan; Rath, Terry and Hill, Nigel 1999. The Use of Value Added Information in Judging School Performance. Institute of Education, University of London. Available at: http://www.cmm.bristol.ac.uk/team/HG_Personal/Full%20Publications%20-%20download/Table%20of%20publications.htm#1997

Grissmer, David 2002. Comment. In Diane Ravitch (ed.), *Brookings Papers in Education Policy 2002*, Brookings Institution Press, Washington DC: 269-272.

Hill, Richard K. and DePascale, Charles A. 2003. Reliability of No Child Left Behind Accountability Designs. *Educational Measurement: Issues and Practice*, 22 (3): 12-20.

Kane, Thomas J. and Staiger, Douglas O. 2002a. Volatility in School Test Scores: Implications for Test-Based Accountability Systems. In Diane Ravitch (ed.), *Brookings Papers in Education Policy 2002*, Brookings Institution Press, Washington DC: 235-283.

Kane, Thomas J. and Staiger, Douglas O. 2002b. The Promise and Pitfalls of Using Imprecise School Accountability Measures. *Journal of Economic Perspectives*, 16 (4): 91-114.

Kane, Thomas J.; Staiger, Douglas O. and Geppert, Jeffrey 2002. Randomly Accountable. *Education Next*, Spring.

Koretz, Daniel 2008. *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press, Cambridge, Mass.

Leckie, George and Goldstein, Harvey 2009. The Limitations of Using School League Tables to Inform School Choice. *Journal of the Royal Statistical Society* Series A, 172

(forthcoming). Available at: http://www.cmm.bristol.ac.uk/team/HG_Personal/Full%20Publications%20-%20download/Table%20of%20publications.htm

Linn, Robert L. 2001a. The Design and Evaluation of Educational Assessment and Accountability Systems. CSE Technical Report 539, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, April. Available at: http://www.cse.ucla.edu/products/rsearch.asp

Linn, Robert L. 2001b. Reporting School Quality in Standards-Based Accountability Systems. Policy Brief 3, National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, Spring. Available at: http://www.cse.ucla.edu/products/policy.html

Linn, Robert L. and Haug, Carolyn 2002. Stability of School-building Accountability Scores and Gains. *Educational Evaluation and Policy Analysis, 24*(1): 29–36.

Newton, Paul E. 2005. The Public Understanding of Measurement Inaccuracy. *British Educational Research Journal*, 31 (4): 419-442.

O'Donoghue, Cathal; Thomas, Sally; Goldstein, Harvey and Knight, Trevor 1997. 1996 DfEE Study of Value Added for 16-18 Year Olds in England. A Report for the Department of Education. Available at: http://www.cmm.bristol.ac.uk/team/HG_Personal/Full%20Publications%20-%20download/Table%20of%20publications.htm#1997

Plewis, Ian and Goldstein, Harvey 1998. Excellence in Schools: A Failure of Standards. *British Journal of Curriculum and Assessment*. 8 (1): 17-20.

Rowe, Ken 2004. Analysing and Reporting Performance Indicator Data: 'Caress' the Data and User Beware! Paper presented at the Public Sector Performance and Reporting Conference, Sydney, April. Available at: http://www.acer.edu.au/documents/Rowe-IIR_Conf_2004_Paper.pdf

Rudd, Kevin 2008a. Quality Education: The Case for an Education Revolution in Our Schools. Address to the National Press Club, Canberra, 27 August. Available at: http://www.pm.gov.au/media/Speech/2008/speech_0443.cfm

Rudd, Kevin 2008b. Questions and Answers. National Press Club, Canberra, 27 August. Available at: http://www.pm.gov.au/media/Interview/2008/interview_0445.cfm

Statistics Commission (UK) 2004. Value Added Measures in School Performance Tables. May. Available at: http://www.statscom.org.uk/C_144.aspx

Wilson, Deborah and Piebalga, Anete 2008. Accurate Performance Measure but Meaningless Ranking Exercise? An Analysis of the English School League Tables. Working Paper No. 07/176, The Centre for Market and Public Organisation, University of Bristol. Available at: http://www.bristol.ac.uk/cmpo/publications/papers/

Zvoch, Keith and Stevens, Joseph J. 2006. Successive Student Cohorts and Longitudinal Growth Models: An Investigation of Elementary School Mathematics Performance. *Education Policy Analysis Archives*, 14 (2). Available at: http://epaa.asu.edu/epaa/v14n2/