
Paper 6: School Performance: Single-level and multilevel analyses of Australian State/Territory comparisons of students' achievements in international studies

Ken Rowe

*Australian Council for Educational Research*¹

Paper presented at the ACSPRI Social Science Methodology Conference
The University of Sydney, 11-13 December 2006

Abstract: Policy activities related to outcomes and standards-based educational performance indicators and their links with growing demands for *accountability*, *standards monitoring*, *benchmarking*, *school effectiveness* and *reform* are widespread and well established in many countries throughout the world. While the long-term goals of school education may be expressed as the enhancement of young peoples' access to and participation in society, as well as preparation for meeting the constantly changing demands of the modern workplace, the most direct and readily accessible measures of student and school performance are obtained from assessments of students' academic achievements. Despite several limitations, achievement data obtained from international studies have several benefits that include: (a) the potential to provide information concerning student and school performance compared with other national systems, and (b) generate understandings (as well as raise questions) about observed differences among educational jurisdictions – within and between countries. However, the collection, analysis and subsequent reporting of such performance indicator data require considerable care. To this end, and for illustrative purposes, the present paper presents findings from fitting both single and multilevel models to students' achievement data for *Reading Literacy* – obtained from participation in the 2003 OECD *Programme of International Student Achievement* (PISA), and compares these achievements at the student and school levels, and between Australia's eight States and Territories. Implications of the findings are discussed.

Introductory comments

The provision of schooling is one of the most massive and ubiquitous undertakings of the modern state. Schools account for substantial proportions of public and private expenditure, and are universally regarded as vital instruments of social and economic policy aimed at promoting individual fulfilment, social progress and national prosperity. It has long been recognised that the key to such prosperity at both the individual and national level is the provision of quality schooling. Since schooling generates a substantial quantity of paid employment for teachers and administrators, it is not surprising that there has long been an interest in knowing how *effective* the provision of school education is and how it can be improved.

The global economic, technological and social changes under way, requiring responses from an increasingly skilled workforce, make high quality schooling an imperative. Whereas OECD education ministers have recently committed their countries to the goal of raising the quality of learning for all, this ambitious goal will not be achieved unless all children, irrespective of their 'intake' characteristics, including backgrounds and locations, receive high-quality schooling and teaching in particular (OECD, 2005a,b).

Central to the goal of providing quality schooling has been the development, construction and increasing use of educational *performance indicators*. Despite several substantive and methodological limitations highlighted by Rowe and Ingvarson (in press), the work of Professor Eric Hanusheck and colleagues at Stanford University (USA) continues to make noteworthy

¹ Correspondence related to this paper should be directed to: Dr Ken Rowe, Research Director, Learning Processes and Contexts Research Program, ACER, 347 Camberwell Road (Private Bag 55), Camberwell, VIC 3142, Australia; *Tel:* +61 3 9835 7489; *Email:* rowek@acer.edu.au.

contributions to understandings about *economic indicators of quality teaching and schooling* (e.g., Hanusheck, 2004, 2005a,b,c; Hanusheck *et al.*, 2005; Hanusheck & Jorgenson, 1996; Hanusheck & Raymond, 2004; Hanushek, Rivkin & Kain, 2005). At this point, a brief outline of what is entailed in such indicators is helpful.

Performance indicators

In general, *performance indicators* (PIs) are defined as *data indices of information* by which the functional *quality* of institutions or systems may be measured and evaluated. Typically, within the context of specified goals and objectives, PI data are ‘measures’ of operational and functional aspects of organizations and/or systems, and provide evidential bases for determining the extent to which such goals and objectives have and are being met. PIs serve various purposes, the most notable of which are for **monitoring, feedback, policy formulation, target-setting, evaluating and reforming**. Although the essential features of educational PIs are consistent with their counterparts in other government and corporate enterprises, they also have unique characteristics – key aspects of which have been highlighted by Rowe (2001a,b, 2004a, 2005a) and by Rowe and Lievesley (2002). At the outset, however, it is helpful to note the importance of educational PIs in prevailing local and international policy contexts.

The nature and purpose of educational PIs

During the last forty years, education systems throughout the world have been subject to considerable reform and change – all justified on the grounds (or at least the rhetoric) of **improving** the *quality* of school education. A key feature of this change has been the frequent revisions of style and policy focus, especially in respect of PIs, with major emphases being placed on the assessment and monitoring of student learning outcomes – mostly in Literacy, Numeracy (Mathematics) and Science. Indeed, current policy activities related to outcomes and standards-based educational PIs and their links with growing demands for *accountability, standards monitoring, benchmarking, school effectiveness* and *reform* are widespread and well established in many developed countries (e.g., Buckingham, 2003; Chapman *et al.*, 1991; Dorn, 1998; Forster, Masters & Rowe, 2001; Hill & Crévola, 1999; Masters, 1990, 1991, 1994, 2004; Rowe, 2001a, 2005a; Tucker & Codding, 1998; Visscher & Coe, 2002; Willms, 2000).

Such emphases are aptly illustrated in the reported proceedings of a meeting under the auspices of the *Summit of the Americas* (2002), which states:

Although it is now part of daily life in schools and in debates between specialists, education assessment has recently become a relevant topic for governments and society, especially because of the economic crisis and the acceleration of the globalization process, which made investments in education a strategic point while the resources available for the sector have shrunk.

In many developed countries including Australia, much of this activity has been (and continues to be) focussed on linking *inputs* and *processes* of educational systems (e.g., physical resources and curriculum provision) with *outputs* (e.g., improvements in student achievement outcomes, as well as in school and system performance). A major effect of such activity has been to signal government policy intention to:

- encourage system accountability to ensure both efficient and effective utilization of resources, and
- bring the delivery of educational services into public sector accounting, underscored by a concern to ensure that such services represent ‘value for money’.

Whereas the provision of quality education is critical to the development of all countries, it is especially the case for developing countries where there is considerable pressure to increase access to education, but not at the expense of quality. Hence, the demand is to ensure that PIs (including assessments of students’ learning and achievements) do not provide a partial, and thus potentially misleading picture of either *quality* or *effectiveness*, as has often been the case.

In spite of difficulties entailed in defining *educational effectiveness* at the school, system, national and international levels, and reaching consensus on the relevant criteria, a good deal of discussion has focused on what is meant by *quality schooling* and *quality teaching*, and how they might be measured and improved. Although the term *quality* is likewise problematic, the "...measurement of the *quality* of schooling is of critical importance at a time when so much school reform in so many parts of the world is being undertaken" (Mortimore, 1991, p. 214). In fact, concerns about the *quality* of school education and its monitoring have long been high priority policy issues in all OECD countries (OECD, 1983, 1986, 1989, 1993, 1995, 2001, 2005a,b). An illustration of this priority is evident in the assertion by Manno (1994):

When judging educational quality, either we focus on what schools spend – or one of its many variants – or we focus on what students achieve, what they know and can do. Those who advocate a focus on outcomes in judging educational quality hold one common belief: we must specify what we expect all children to learn, and we must assess them to determine whether they have learned it.

While the long-term goals of school education may be expressed as the enhancement of young peoples' access to and participation in society, as well as preparation for meeting the constantly changing demands of the modern workplace (OECD, 1983, 1986, 2005a), the most direct and readily accessible measures of schooling outcomes are obtained from assessments of students' academic attainments. Herein, however, lies a dilemma that is evidenced in strident critiques of traditional and prevailing psychometric models for test and examination modes of assessment (e.g., Berlak, 1992) and an equally strident chorus of concern for the deleterious effects of test-driven and 'test-dominated' curricula (e.g., Kellaghan, Madaus & Airasian, 1992; Lacey & Lawton, 1981). As Watson (1996) noted: "In high stakes testing environments, educational practitioners are likely to distort their behaviour in order to meet the demands of the indicator, usually to the detriment of their real job" (p. 119). Nisbet (1993, p. 25) further highlighted this dilemma in the following terms:

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

Although measures of student learning and achievement outcomes are prime PIs of education systems and the services they provide and for which they are responsible, there are many others (including both *inputs* and *processes*) that constitute useful bases for informed planning and decision-making, followed by implementation and reform. If decisions for improvement are to be **data-informed**, rather than based on whim or ideology, then useful, dependable and timely information on PIs is required. Indeed, such bases constitute key purposes of specifying, gathering and using PIs for educational change and reform. In particular, PI information allows systems and their constituent organizational elements to: (1) formulate strategic policy priorities and their related targets, (2) specify achievable objectives, (3) implement them, and (4) evaluate the extent to which those target objectives have been attained.

The benefits and limitations of national and international assessment programs for monitoring student learning and achievement outcomes

The benefits and limitations of national and international monitoring programs for student learning and achievement outcomes have been well documented and require little reiteration here (e.g., Beaton *et al.*, 1999; Forster, 2000, 2001a,b; Goldstein, 2001, 2004; Greaney & Kellaghan, 1996; McGaw, 1991; Murphy *et al.*, 1996; Plomp, 1999; Rowe, 2004a; Rowe & Lievesley, 2002; Scheerens & Bosker, 1997; Visscher *et al.*, 2000; Willms, 2000). In brief, the benefits of national assessments include the provision of systematic and regular measures of student learning and achievement outcomes. They are designed to evaluate the relative 'health'

of education systems, to monitor achievement across the systems, and provide information that allow comparisons of performance within the system of sub groups of students, as well as within and between districts, regions and states. The data obtained assist policy makers to allocate resources designed to maximize learning opportunities and outcomes for all students. Nonetheless, McGaw (1991, p. 138) pointed out both the benefits and risks involved in national achievement monitoring programs in the following terms:

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

Subsequently, and consistent with the warnings of Goldstein and Spiegelhalter (1996) about the dangers of publishing student and school performance data in the form of 'league tables', Rowe (2000, p. 92) observed:

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

While there are distinct advantages in implementing assessment programs at the beginning of the school year for *diagnostic* purposes to assist teachers in meeting the specific learning needs of students (both at the individual and cohort levels) as in France (see OECD, 1993), accountability pressures on State and Federal governments in Australia to monitor educational standards are political realities, and ones that are not likely to diminish. In this context, Hill (1995, p. 4) noted:

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

Achievement data obtained from both national and international studies have several benefits. In the case of international studies, given that the measurements of students' achievements are calibrated on common scales, such benefits include: (a) the potential to provide valuable PI information about a country's education system(s) in relation to other national systems concerning the performance of students and schools, and (b) generate understandings (as well as raise questions) about observed differences in the achievements of students from different educational systems. For example, Plomp (1999, pp. 1-2) has noted:

The understandings we obtain from cross-national comparisons of such policies as age of school entry, hours and methods of instruction, and teacher training, can provide us with new insights into the performance of our own educational system in general, and of the relationship between student performance and its antecedents and consequences in particular.

Findings from international studies of student achievement also have the advantage of attracting political and media attention. Thus, poor results can provide policy makers with a strategic rationale for intervention and budgetary support advocacy throughout education systems and their constituent jurisdictions (see: Forster, 2000, 2001a,b; Greaney & Kellaghan, 1996; Rowe & Lievesley, 2002). However, several studies have now shown that there are serious and *inherent* limitations to the usefulness of such indicators for providing reliable judgements about educational institutions (e.g., Goldstein & Thomas, 1996; Goldstein &

Spiegelhalter, 1996; Marsh, Rowe & Martin, 2002; Rowe, 2000, 2004a; Visscher *et al.*, 2000). Key reasons for these limitations are as follows:

- Against the background of what is known about *differential school effectiveness* (e.g., Nuttall, Goldstein, Prosser & Rasbash, 1989) it is not possible to provide simple summaries that capture all the important features of schools (see also: Bosker, Creemers & Scheerens, 1994; Hill & Rowe, 1996, 1998; Rowe, 2000, 2004a; Rowe & Hill, 1998; Rowe & Rowe, 1999; Rowe, Turner & Lane, 2002; Rowe & Stephanou, 2001, 2003; Scheerens & Bosker, 1997; Stephanou & Rowe, 2002; Visscher *et al.*, 2000; Willms, 2000).
- By the time information from a particular school has been analysed, it refers to a 'cohort' of students who entered that school several years previously so that its usefulness for *future* students and the making of judgements about *school effectiveness* may well be dubious. Where information is analysed on a yearly basis, it is necessary to make adjustments for prior contributing factors that extend over two or more years. In fact, it is increasingly recognised that schools, or teachers within those schools, should not be judged by a single 'cohort' of students, but rather on their performance over time (e.g., Goldstein, 1997; Thomas *et al.*, 1997; Thomson *et al.*, 2005). As noted by Goldstein (1997), this makes the historical nature of *school effectiveness* judgements an acute problem.
- Above all, even when suitable adjustments for students' intake characteristics and prior achievements have been taken into account, the resulting *value-added* estimates have too much *uncertainty* attached to them to provide reliable rankings. This point, illustrated elsewhere, is vital and one that is all too-frequently ignored by advocates of published 'league tables' (see: Rowe, 2000, 2004a; Rowe & Stephanou, 2001, 2003; Rowe, Turner & Lane, 2002).

Limitations of findings summarised in this paper

The limitations of findings from analyses of data derived from national and international assessment programs for monitoring student learning and achievement outcomes, as outlined above, apply equally to those presented in this paper. Principal among these is the limited number of available explanatory variables at the student-level and group-membership levels (e.g., class, school and State/Territory) to provide effect estimates required for adjustment. For example, the fact that Australian students (and their parents) are not obliged to disclose their ethnic (and religious) affiliations in surveys of any kind results in large proportions of 'missing data' for these variables.

This is particularly relevant to obtaining effect estimates for students' Indigenous status (i.e., Aboriginal and Torres Strait Island 'membership', or otherwise). Indeed, there is always the difficulty that any statistical model used to provide effect estimates at the student, contextual and group-membership levels, will fail to incorporate **all** the appropriate adjustments, or in some other way may be mis-specified. Thus, at best, effect estimates can only be used as 'screening devices' to identify 'outliers' (which could form the basis for follow-up), but they cannot and should not be used as definitive measures of the effect of those schools (or jurisdictions) on student learning *per se*.

A further limitation that applies to the present paper is the restriction on providing cross-sectoral comparisons of student achievement outcomes (i.e., across government, Catholic and independent schools). This restriction derives from directives by national Steering Committees for non-disclosure of cross-sectoral comparisons of findings from both national monitoring programs and international studies. Regardless of justifications for these directives, it is important to note that the results presented here lack adjustments for this major source of contextual variation – particularly given the increasing student enrolment 'drift' from government to non-government schools during the last decade in all Australian States and

Territories since 2000.² This deficiency imposes severe restrictions on estimating the differential effects of teaching and learning provision within and between sectors (see: Cuttance, 2001; Darling-Hammond, 2000; Hanushek, Rivkin & Kain, 2005; Hill & Rowe, 1996; Muijs & Reynolds, 2001; Rowe, 2004b).

Focus of the present paper

Despite the importance of the substantive issues noted above, the focus of the present paper is primarily methodological. Whereas substantive and methodological issues related to analyses of PI data are not mutually exclusive, all too often analysts of PI data fail to account for the *measurement, distributional* and *structural* properties of the obtained data – particularly in fitting single-level explanatory models under ordinary least squares (OLS) estimation, even though such data are inherently hierarchical. In consequence, magnitude estimates of major sources of contextual and/or structural variance (including error variance in the measurement of both response and explanatory variables) are not accounted for.³ Such basic methodological deficiencies invariably yield misestimates of parameters and their standard errors that frequently lead to misleading interpretations of the findings. Hence, for illustrative purposes, this paper presents findings from fitting both single- and multilevel explanatory models to student achievement data for a designed sample of 15-year-old students (typically in their eleventh year of schooling) located in schools throughout Australia's six States and two Territories.

The data

The data analysed and reported here have been obtained from Australia's participation the 2003 OECD *Programme of International Student Assessment (PISA)* with a focus on student performance in *Reading Literacy*.⁴ In particular, the data derive from a stratified sample of 12,551 15-year-old students, drawn from 321 government, Catholic and independent schools located within six Australian States and two Territories. For detailed documentation related to Australia's participation in the 2003 PISA study, see Thomson, Cresswell and De Bortoli (2004).

Due to variations across Australia's States and Territories in respect of school starting ages (and hence, age/grade membership), adjustments are made for participating students' *Age* (in years and months) and *Grade* (Grade level, or the number of years of formal schooling). Whereas the largest source of variation in school performance is typically attributed to differences in what students bring to school, including: their abilities and attitudes, family and community wealth and background, the research evidence shows that school systems within Australia differ in the extent to which students' intake characteristics and socioeconomic (SES) background influences achievement (Marks, 2000, 2005, 2006; OECD, 2002). The related research findings also show that there is often substantial variation in student performance within- and between-schools serving similar socioeconomic catchment areas, as well as between classes within the same school (e.g., Hill & Rowe, 1996, 1998; Marks, 2005, 2006; Rowe, 2001b, 2003, 2004a). These differences imply that school system policies, and individual school

² See ABS (2006) for evidence of this 'drift' during the period 1980-2005.

³ A key reason for this is that in fitting single-level *general linear models* to PI data under OLS estimation, both the response and explanatory variables are assumed to be 'measured' without error. For an explication of the inappropriateness of this assumption, see Rowe (1989, 2004a).

⁴ In the OECD *Programme for International Student Assessment (PISA)*, the construct of *Reading Literacy* emphasises skill in using written information in situations that students may encounter in their life both at and beyond school. Thus, *Reading Literacy* is defined as: '*... understanding, using and reflecting on written texts in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society*' (OECD, 2003, p. 108). For specific details related to the PISA 2000 and 2003 results relevant to Australia, see: Lokan, Greenwood and Cresswell (2001); Thomson, Cresswell and De Bortoli (2004). For comprehensive analyses of Australian students' measured achievements in PISA 2000 and 2003 for *Reading Literacy, Mathematical Literacy* and *Scientific Literacy*, see Rowe (2006a).

and teacher practices, do make a difference in influencing student learning and achievement outcomes. Thus, adjustments are also made for student *Gender* (coded '1' for females, '0' for males), and for the continuous variables of *SES* and *Home Educational Resources* (HEDRES). For specific details of the measurement properties and definitions for both achievement and presage variables relevant to the PISA studies, see: Lokan, Greenwood and Cresswell (2001); Thomson, Cresswell and De Bortoli (2004).

Findings from fitting single-level models to the data

Following specifications of the fitted models, for ease of reporting and interpretation, the findings are presented graphically.⁵

To account for variation in students' *Reading Literacy* achievement scores, a typical, single-level multivariate regression model for these data may be specified as follows:

$$\text{Reading}_i = \beta_0 + \beta_1 \text{AGE}_i + \beta_2 \text{Grade}_i + \beta_3 \text{Gender}_i + \beta_4 \text{SES}_i + \beta_5 \text{HEDRES}_i + \beta_6 \text{ScAvSES}_i + e_i$$

where *Reading* is the response variable – being the *Reading Literacy* achievement score for student *i*, β_0 is the intercept term, and AGE, Grade, SES, HEDRES and ScAvSES⁶ are the fitted explanatory variables. The term e_i is the student-level prediction residual – the proportion of variance in which is that proportion of variance in the *Reading* response variable that is unaccounted for after fitting the explanatory variables. Note that since the response and explanatory variables are 'measured' on different metrics, for subsequent interpretation purposes when fitting explanatory regression models to such data, it is important normalize the raw data – preferably as rank-ordered, normal-equivalent-deviates (NEDs) under the Normal distribution. The magnitudes of the obtained parameter estimates from the fitted model may then be interpreted in terms of standard deviation (SD) effect sizes. Under an OLS method of estimation, the solution to the fitted single-level model specified above is given in Table 1.

Table 1. Solution to Fitted Single-level Model*

Variable/Term	Parameter Estimate (SD units)	SE	t-Value
Intercept (β_0)	-0.156	0.011	-14.18
AGE (β_1)	-0.048	0.009	-5.11
Grade (β_2)	0.262	0.012	22.42
Gender (β_3)	0.359	0.016	22.35
SES (β_4)	0.081	0.009	9.16
HEDRES (β_5)	0.250	0.011	23.55
ScAvSES (β_6)	0.214	0.009	24.67
Residual (e_i)	0.804	0.010	80.40

* Note: all recorded parameter estimates are statistically significant beyond the $p < 0.05$ α level by univariate 2-tailed test.

⁵ Due to their efficiency and high-resolution graphics capabilities, the data analyses, explanatory modelling and graphical presentations of the findings were undertaken using *STATISTICA* (StatSoft, 2005) and *MLwiN* (Browne, Healy, Cameron & Charlton, 2005).

⁶ Note that ScAvSES (*School Average SES*) is also included here to account for the school cohort effect of Socio-economic Status (SES), over and above that operating at the individual student SES level. Whereas this variable is strictly a level-2 variable, it should be noted that it is subscripted by *i* such that in the single-level case, students within each school are assigned the same ScAvSES score.

Comment: From the findings presented in Table 1, around the normalised grand mean for students' *Reading Literacy* scores of -0.156 units, the fitted variables accounted for a mere 19.3% of the variance in *Reading Literacy* scores, with a large and significant residual. While the effect of AGE was significantly negative (i.e., in favour of younger students), each of the other fitted explanatory variables were significantly positive, including: Grade (in favour of students in higher Grade levels); Gender (in favour of females), SES (in favour of higher SES); HEDRES (in favour of students with higher levels of *Home Educational Resources*); and ScAvSES (in favour of students within schools with higher *School Average SES*).

While these findings are of some interest, they are not especially informative – particularly in terms of interpreting student Gender differences across the eight State/Territory systems, for example. To this end, the findings from a further single-level, multivariate analysis of variance (MANOVA) model to the relevant data are helpful – adjusted for the main effects of *State/Territory* location, *Gender*, *Age*, *Grade*, *SES*, and *Home Educational Resources* (HEDRES). The findings from the fitted model are summarised in Figure 1.

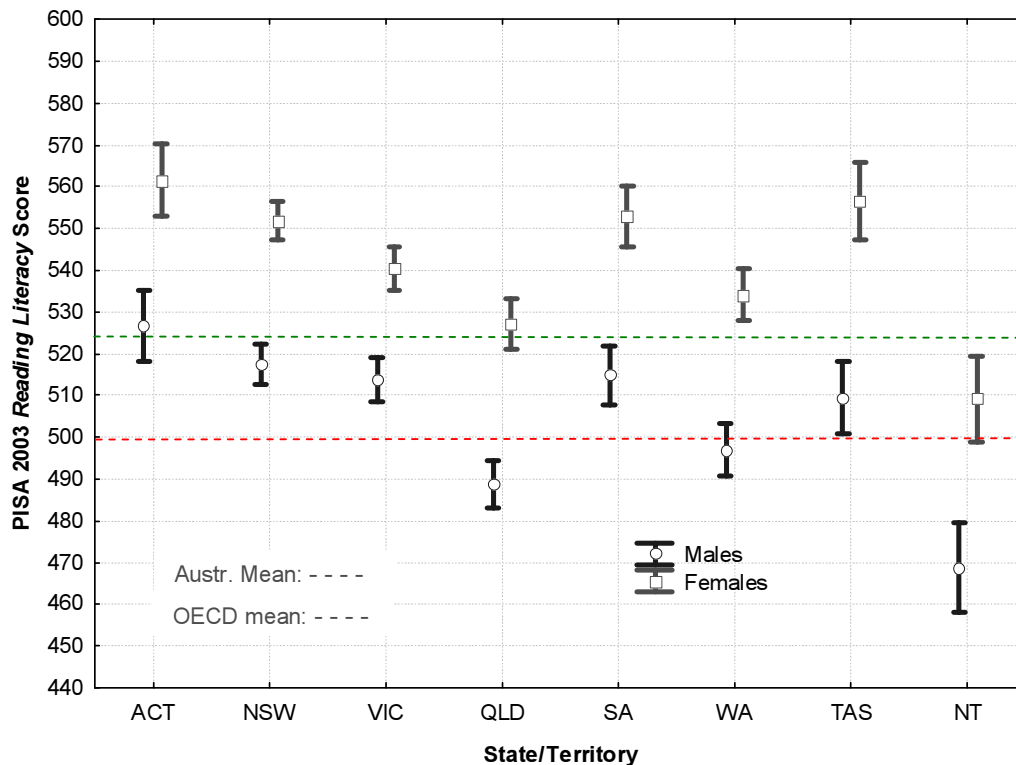


Figure 1. Adjusted mean-point estimates of students' PISA 2003 Reading Literacy scores bounded by 95% confidence intervals, by State/Territory and Gender

State/Territory effect: $F(7,12481) = 33.1, p < 0.000001$
 Gender effect: $F(1,12481) = 394.9, p < 0.000001$
 Age effect: $F(1,12481) = 63.3, p < 0.000001$
 Grade effect: $F(1,12481) = 540.8, p < 0.000001$
 SES effect: $F(1,12481) = 221.9, p < 0.000001$
 HEDRES effect: $F(1,12481) = 763.5, p < 0.000001$
 State/Territory \times Gender effect: $F(7,12481) = 1.5, p = 0.179$ (n.s.)

Comment: With the exception of the *State/Territory* \times *Gender* interaction effect, all main effects were statistically significant, the most notable of which were: *Home Educational Resources* (HEDRES – in favour of higher HEDRES), *Grade* (in favour of higher Grade membership), *Gender* (in favour of females), *SES* (in favour of higher SES), and *Age* (in favour

of younger students). A comparison of Gender differences from the PISA 2000 and 2003 data for students' achievements for *Reading Literacy* is of interest here.

Compared with the equivalent findings for PISA 2000 *Reading Literacy* (see Rowe, 2006a, Fig. A2.1), the 2003 findings indicated a wider 'gap' between the performance levels of females and males (in favour of females) in all States and Territories, including among ACT students where the 'gender gap' was not statistically significant in 2000. The means for males in QLD and NT schools were significantly below the OECD average. This increasing achievement 'gap' in relative under-achievement of males is of both policy and practice concern – as evident in the raw, unadjusted mean score data summarised in Table 2.

Table 2. Female-Male Unadjusted Mean Score Differences for PISA 2000 and 2003 Reading Literacy, by Australian States and Territories¹

Mean Female-Male Difference	ACT	NSW	VIC	QLD	SA	WA	TAS	NT
² PISA 2003	42	39	30	49	34	40	45	58
³ PISA 2000	23	30	28	47	29	34	50	30

¹ Adapted from Thomson, Creswell and De Bortoli (2004, p. 105)

² All female-male differences are statistically significant at the $p < 0.05$ level.

³ All female-male differences are statistically significant at the $p < 0.05$ level, except for ACT

Whereas these results are informative, it is important that findings from fitting single-level models to the student achievement PI data are not over-interpreted, since such models ignore the inherent hierarchical structure of the data; i.e., 12,551 students' achievement scores (level-1) clustered within 321 schools (level-2) and 8 States/Territories (level-3). This structure requires that multilevel models be fitted to the data.

Findings from fitting multilevel models to the data

Following are results of the fitted baseline Variance Components model for 12,551 15 year-old students (i) in 321 schools (j) drawn from 8 Australian States/Territories (k) – based on students' normalised scaled scores for PISA 2003 *Reading Literacy*. Using an iterative generalized least squares (IGLS) method of estimation (see Goldstein, 1986), the results present the normalised parameter estimates (coloured green) and their standard errors in parentheses (also coloured green) for the **residual** variation (res. var.) at: the State/Territory-level (v_{0k}), between-school-level (u_{0jk}), and within-school (student)-level (e_{0ijk}).⁷

$$\text{Reading}_{ijk} = \beta_{0ijk} \text{Cons}$$

$$\beta_{0ijk} = -0.032(0.056) + v_{0k} + u_{0jk} + e_{0ijk}$$

$$\begin{bmatrix} v_{0k} \end{bmatrix} \sim N(0, \Omega_v) : \Omega_v = \begin{bmatrix} 0.018(0.012) \end{bmatrix} \text{ Between-State/Territory res. var: 1.9\% (n.s.)}$$

$$\begin{bmatrix} u_{0jk} \end{bmatrix} \sim N(0, \Omega_u) : \Omega_u = \begin{bmatrix} 0.196(0.017) \end{bmatrix} \text{ Between-schools res. var: 20.8\%}$$

$$\begin{bmatrix} e_{0ijk} \end{bmatrix} \sim N(0, \Omega_e) : \Omega_e = \begin{bmatrix} 0.792(0.010) \end{bmatrix} \text{ Within-schools res. var: 77.3\%}$$

Comment: Around the normalised grand mean for Australia (-0.032), the residual variance in students' PISA 2003 *Reading Literacy* achievement scores between States and Territories was

⁷ Note that when the magnitude of a parameter estimate is at least twice its corresponding standard error, statistical significance is indicated at and beyond the $p < 0.05$ α level.

not statistically significant (with the possible exception of NT). Following Goldstein and Healy (1995), the unadjusted residual plots at the State/Territory-level illustrate these results as follows:

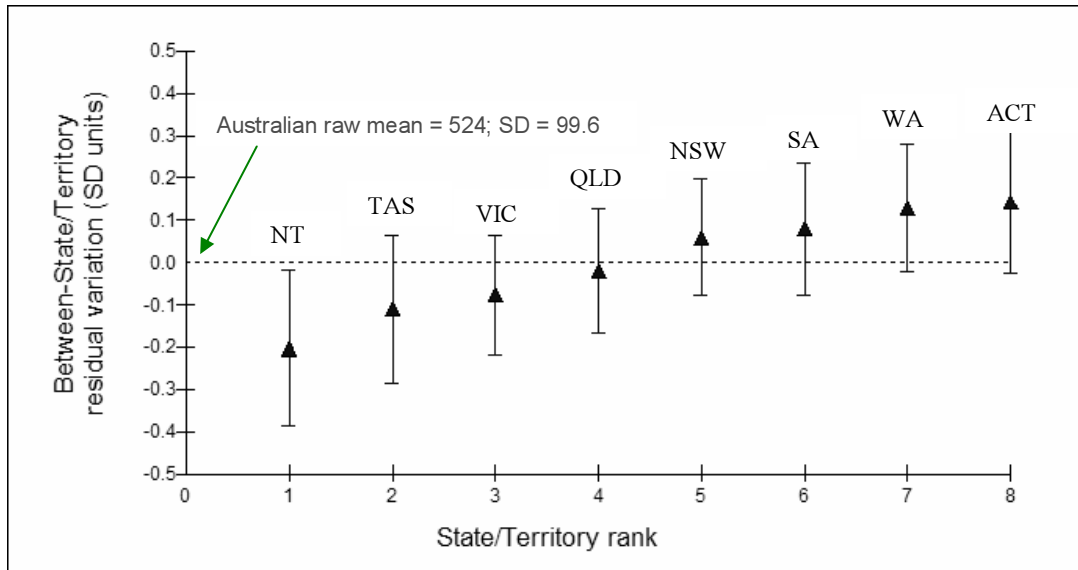


Figure 2. Ranked State/Territory-level raw residuals for PISA 2003 Reading Literacy scores, showing mean-point estimates, bounded by 95% ‘uncertainty’ intervals

Comment: Apart from NT (below), the ‘uncertainty’ intervals around the unadjusted means for other States and Territories all overlap the ‘population’ mean (zero) – indicating non-significant differences at the State/Territory-level in students’ PISA 2003 Reading Literacy scores – accounting for a mere 1.9% of the residual variance (cf. findings from fitting the single-level MANOVA model summarised in Figure 1 above). Although the estimates obtained from fitting this multilevel variance-components model to the data are not of particular interest (*per se*), they provide a useful base from which to compare findings from fitting more ‘responsible’ models.

In the following model, adjustments are made for the ‘intake’ variables of *Gender*, *Age*, *Grade*, family *SES* (at the student-level), *Home Educational Resources* (HEDRES), and school average *SES* at the school-level (i.e., *ScAvSES* – to estimate the within-school average ‘cohort effect’ of SES, over-and-above that operating at the individual student-level). The results of the fitted model to the normalized data are given below, indicating the magnitude of the parameter estimates for the fitted variables (in SD units), and their respective standard errors given in parentheses. [Note again that statistical significance at the $p < 0.05$ level is indicated when parameter estimates are at least twice the magnitude of their corresponding standard errors].

$$\text{Reading}_{ijk} = \beta_{0ijk}\text{Cons} + 0.352(0.017)\text{Gender}_{ijk} + -0.069(0.009)\text{AGE}_{ijk} + 0.310(0.013)\text{Grade}_{ijk} + 0.076(0.008)\text{SES}_{ijk} + 0.192(0.010)\text{HEDRES}_{ijk} + 0.208(0.019)\text{ScAvSES}_{jk}$$

$$\beta_{0ijk} = -0.194(0.057) + v_{0k} + u_{0jk} + e_{0ijk}$$

$$[v_{0k}] \sim N(0, \Omega_v) : \Omega_v = [0.022(0.013)] \text{ Between-State/Territory res. var: 2.7\% (n.s.)}$$

$$[u_{0jk}] \sim N(0, \Omega_u) : \Omega_u = [0.094(0.009)] \text{ Between-schools res. var: 11.5\%}$$

$$[e_{0ijk}] \sim N(0, \Omega_e) : \Omega_e = [0.703(0.009)] \text{ Within-schools res. var: 85.8\%}$$

Comment: These results indicate significant effects for: *Gender* (in favour of females), *Grade* (in favour of higher Grade membership), both *SES* at the student-level and *ScAvSES* at the school-level, as well as *Home Educational Resources* (HEDRES), were significant predictors of Australian students' PISA 2003 *Reading Literacy* achievement scores. While the effect of *Age* was small, it was statistically significant (in favour of younger students).

Together, all six fitted variables accounted for only 13.2% of the variance in students' achievement scores, with an insignificant 2.7% of the residual variance at the State/Territory-level, and a significant 11.5% of the residual variance due to variation between schools. As expected, the bulk of the residual variance was at the student-level (i.e., 85.8%). Mean-adjusted residual plots at the State/Territory-level illustrate these results are presented in Figure 3 below.

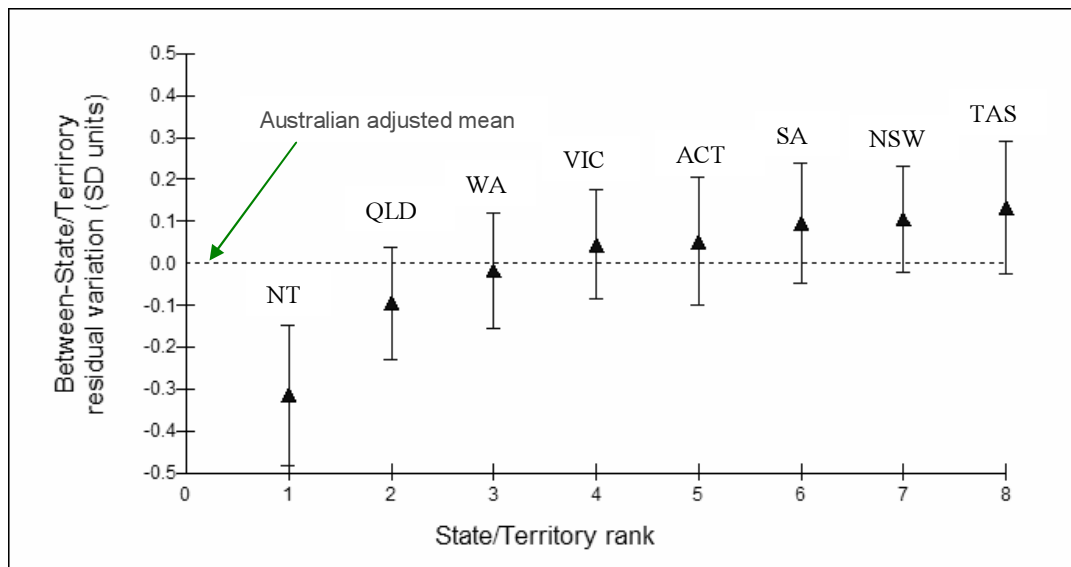


Figure 3. Plot of ranked State/Territory residuals, showing adjusted mean-point PISA 2003 *Reading Literacy* score estimates, bounded by 95% 'uncertainty' intervals

Comment: With the exception of NT (below), the 95% 'uncertainty' intervals around the adjusted means, the intervals for the other States and Territories all overlap the 'population' mean (zero) – indicating non-significant differences at the State/Territory-level (i.e., a small 2.7% of the residual variance in students' PISA 2003 *Reading Literacy* scores). Indeed, there were no significant differences between the adjusted mean performances of students located in QLD, WA, VIC, ACT, SA, NSW, and TAS schools.

Although these results are of minor interest, they mask the variation between-schools at the national level, as well as between-school variation within each of the States and Territories separately. Following in Figure 4 is a plot of ranked residuals for 321 Australian schools from a multilevel analysis of residuals for performance in PISA 2003 *Reading Literacy*, showing adjusted mean-point estimates bounded by 95% 'uncertainty' intervals.

Comment: From Figure 4, the 'uncertainty' intervals around the adjusted means for approximately 85% the 321 schools all overlap the national mean (zero) – indicating non-significant differences in students' *Reading Literacy* scores between these schools. The remaining 15% of schools yielded student performances either significantly above or below the national mean, resulting in an increased between-school residual variation (11.5%) compared with PISA 2000 *Reading Literacy* (5.5%; see Rowe, 2006a, Fig. A2.5). These findings again indicated differences in the quality of teaching and learning provision among Australian schools.

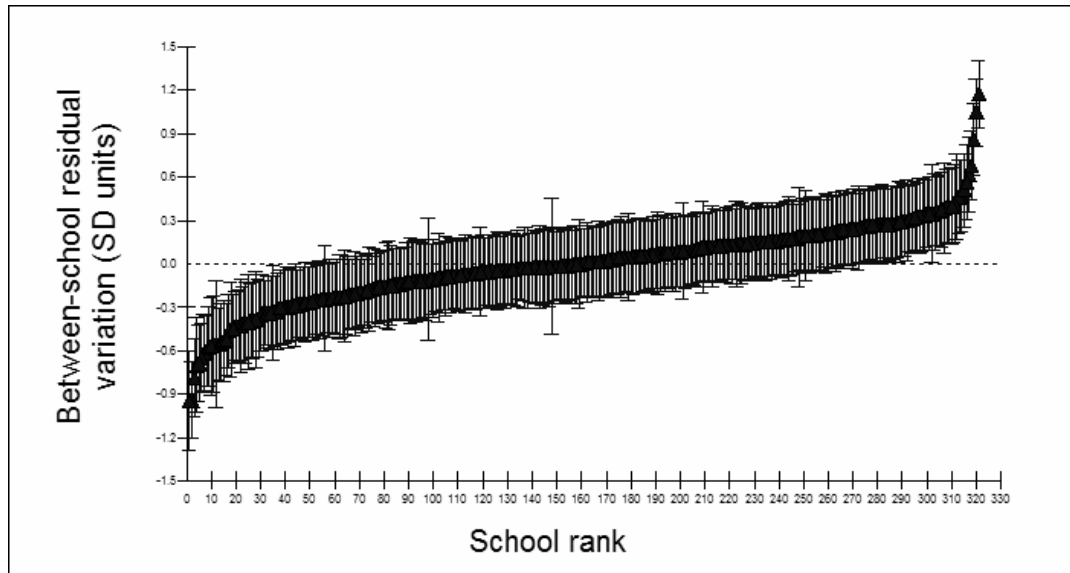


Figure 4. Plot of ranked residuals for 321 schools, showing adjusted mean-point PISA 2003 *Reading Literacy* score estimates, bounded by 95% 'uncertainty' intervals

Discussion

Notwithstanding the limitations in the data analysed and reported here – as noted earlier – the findings have important substantive and methodological implications. Among the substantive implications are that *Reading Literacy* competence constitutes the foundational skill that underlies effective engagement with the school curriculum. This assertion is supported by the work of Nobel Prize winning economist James Heckman's (2000, 2005) overview of the economic aspects of human skills formation. Heckman concludes that investment in the learning development of children and young people is crucial. For Heckman, literacy competence is an essential area of learning investment in the young, being a 'skill that begets many other skills' (an index of 'self-productivity', as he calls it), because it constitutes a 'key part of our capacity to increase our capacity'.

From the findings of fitting single-level and multilevel models to the PISA 2003 data presented here, three substantive and methodological features are worthy of emphasis. First, on average, the achievement performances of Australian 15-year-old students in *Reading Literacy* have consistently been significantly above the OECD averages. As a nation, this result is to be celebrated – reflecting positively on the quality of teaching and learning provision in Australian schools. Of concern, however, is the increased 'gap' between 2000 and 2003 that separated the mean *Reading Literacy* achievements of male and female students (see Table 2).

Second, although these international assessments of *Reading Literacy* during 2003 indicated that 15-year-old students in Australian schools performed notably better (on average) than the majority of their counterparts in other OECD countries, there have been notable variations between States/Territories and sub-groups of students within them. For example, as reported by Thomson, Cresswell and De Bortoli (2004), 12 per cent of students (ACT, WA) to 28 per cent (NT) had not developed the literacy skills needed for further education, training and work (defined as *low achievers*), particularly indigenous students (35%) and males (17%).

Similar proportional estimates have been reported for achievement in reading comprehension of 14-year-old Australian students between 1975 and 1998, and, with few exceptions, the

estimates have remained constant during the period.⁸ Furthermore, approximately 20 per cent of Australians aged 15-74 years have been identified as having “very poor” literacy skills, with an additional 28 per cent who “could be expected to experience some difficulties in using many of the printed materials that may be encountered in daily life” (ABS, 1997, p. 7). The importance of competence in reading for achievement in science and mathematics has already been noted and illustrated by Rowe (2006a) – low performance in which severely limits opportunities for further education and training, as well as active participation in economic and social life.

Third, the comparative State/Territory findings from fitting single-level explanatory models to the PISA 2003 student achievement data for *Reading Literacy* presented in Figure 1, suggest significant variation between the States and Territories, even following adjustments for students’ background and ‘intake’ characteristics. At the *prima facie* level, these findings indicate that the performances of students in ACT schools compared favourably with their counterparts in the other States and Territories. Indeed, the achievement performances of students in ACT schools for *Reading Literacy* were notably ‘better’ than their counterparts in other States and Territories. Nevertheless, the results from fitting more ‘responsible’ multilevel models to the achievement data (with similar adjustments), clearly indicated that these ‘observed’ differences were inflated and hence, misleading. These outcomes reflect a widespread failure to understand that PI data collected at level-1 (students in this case) are not independent of the contexts in which they are gathered (i.e., schools and States/Territories in the present case).

Regretfully, within Australia (as in many other countries), the clustering effects of students-within schools (and higher level contexts) are all-too frequently ignored by educational PI data analysts. In consequence, the resulting aggregation bias yields misestimates of the effects at best, and misleading findings at worst (Rowe, 2004a). In contrast, findings from multilevel analyses of the data are more ‘responsible’ since they more accurately reflect the ecological reality of students being nested within-classes and schools, etc., such that major sources of variation may be identified and estimated. Thus, the key methodological message of the present paper is that responsible estimation and reporting of these effects are not possible from fitting single-level models to such inherent hierarchically-structured PI data.

Concluding comments

The issues surrounding *school performance* and *educational effectiveness* are complex, multivariate, multidimensional and multilevel. While Australia has much to be proud of its schools and the achievements of students within them, the findings summarised in this paper indicate considerable within and between-school variations, if not between its six States and two Territories. Ultimately, however, *quality schooling* and *educational effectiveness* for all students is crucially dependent on the provision of *quality teaching* by competent teachers who are supported by capacity-building towards the maintenance of high teaching standards via initial quality pre-service education and subsequent in-service professional development at all levels of schooling (see: Darling-Hammond, 2000; Darling-Hammond & Bransford, 2005; Ingvarson, 2002, 2003; 2005; Ingvarson *et al.*, 2006). For obvious reasons, it is vital that this capacity-building is firmly grounded in findings from evidence based research – especially in pedagogical practices that are demonstrably effectively in engendering students’ learning and achievement outcomes – rather than prevailing ideological commitments to ‘strategies’ that have no research basis (see: Hattie, 1987; 2003, 2005; Hattie, Biggs & Purdie, 1996; Hoad *et al.*, 2005; Rowe, 2005b,c, 2006b; Westwood, 2004, 2006).

Such outcomes, however, call for major reform requiring an investment in teacher quality that can then be used to change the ways in which students are taught and learn. Sadly, many

⁸ See Rothman (2002), who notes: “For some groups, there has been improvement, most notably for students from language backgrounds other than English. For other groups, however, results indicate a significant achievement gap. The most significant gap is between Indigenous Australian students and all other students in Australian schools” (p. ix).

educational reforms stop short of changing what happens beyond the classroom door, and thus fail to deliver improved teaching and learning outcomes for teachers and students, respectively. Rather, real reform directed at improving outcomes for all students – regardless of their backgrounds, ‘intake’ characteristics and residential locations – calls for substantial change in the quality of *teaching* and *learning* provision, but unless there is total commitment to teacher capacity-building, reform efforts soon falter.

It is important to note that the ‘myth’ of *school effectiveness* is grounded in a widespread failure to understand the fundamental distinction between *structure* and *function* in school education. Whereas a key *function* of schools is the provision of quality teaching and learning experiences that meet the developmental and learning needs of students is dependent on funding and organisational *structures* that support this function, the danger is a typical proclivity on the part of educational administrators to stress *structure* (e.g., single-sex schooling, class size, curriculum construction and reconstruction, etc.) at the expense of *function* (quality teaching and learning). Unfortunately, such emphases are indicative of a pervasive ignorance about what **really** matters in school education (i.e., quality teaching and learning), and the location of major sources of variation in students’ educational outcomes (i.e., the classroom).

It seems we need to be constantly reminded that schools and their structural arrangements are only as effective as the those responsible for making them work (school leaders and teachers) – in cooperation with those for whom they are charged and obligated to provide a professional service (students and parents). We also need to be reminded that the most valuable resources available to schools and the performance of their students are teachers. Thus, for the sake of Australia’s social and economic future – at the individual, school and national levels – we need to improve student and *school performance* by investing in *teacher* and *teaching quality*. By any criterion, the substantive and methodological implications for irresponsible analyses and reporting of such PI data are untenable for both policy and practice.

References

- ABS (1997). *Aspects of literacy: Assessed literacy skills*. Canberra, ACT: Australian Bureau of Statistics (4228.0). Available at: <http://www.abs.gov.au/Ausstats/abs@.nsf/0/887AE32D628DC922CA2568A900139365?Open>.
- ABS (2006). *Schools Australia 2005* (Cat. No: 4221.0). Canberra, ACT: Australian Bureau of Statistics.
- Beaton, A., Postlethwaite, T., Ross, K., Spearritt, D., & Wolf, R. (1999). *The benefits and limitations of international achievement studies*. Paris: International Institute for Educational Planning/UNESCO.
- Berlak, H. (1992). The need for a new science of assessment. In H. Berlak, F.M. Newmann, E. Adams, D.A. Archbald, T. Burgess, J. Raven, & T.A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 139-180). Albany: State University of New York Press.
- Bosker, R.J., Creemers, B.P.M., & Scheerens, J. (Guest Editors) (1994). Conceptual and methodological advances in educational effectiveness research. *International Journal of Educational Research*, 21(2), 121-231.
- Browne, W., Healy, M., Cameron, B., & Charlton, C. (2005). *MLwiN software for multilevel analysis* (Version 2.02). Bristol, UK: Centre for Multilevel Modelling, University of Bristol. Information and downloads available at: <http://www.emm.bristol.ac.uk>.
- Buckingham, J. (2003). *Schools in the Spotlight: School performance and public accountability* (CIS Policy Monograph 59). Sydney, NSW: The Centre for Independent Studies.
- Chapman, J., Angus, L., Burke, G., & Wilkinson, V. (Eds.) (1991). *Improving the quality of Australian schools*. Australian Education Review No. 33. Hawthorn, VIC: Australian Council for Educational Research.
- Cuttance, P. (2001). The impact of teaching on student learning. In K.J. Kennedy (Ed.), *Beyond the rhetoric: Building a teaching profession to support quality teaching* (pp. 35-55). Deakin West, ACT: Australian College of Education; College Year Book 2001.
- Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives*, 8(1); available at: <http://epaa.asu.edu/epaa/v8n1>.

- Darling-Hammond, L., & Bransford, J. (Eds.) (2005). *Preparing teachers for a changing world: What teachers should learn and be able to do*. San Francisco, CA: Jossey-Bass.
- Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1), 1-32.
- Forster, M. (2000). *A policy maker's guide to international achievement studies*. Camberwell, VIC: Australian Council for Educational Research.
- Forster, M. (2001a). Using assessment data. In M. Forster, G.N. Masters and K.J. Rowe, Measuring learning outcomes: Options and challenges in evaluation and performance monitoring (Revised edition, pp. 38-54). *Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring*. Washington, DC: The World Bank Institute.
- Forster, M. (2001b). *A policy maker's guide to system-wide assessment programs*. Camberwell, VIC: Australian Council for Educational Research.
- Forster, M., Masters, G.N., & Rowe, K.J. (2001). Measuring learning outcomes: Options and challenges in evaluation and performance monitoring. *Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring*. Washington, DC: The World Bank Institute.
- Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 73, 43-56.
- Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, 8, 369-395.
- Goldstein, H. (2001). Using pupil performance data for judging schools and teachers: Scope and limitations. *British Educational Research Journal* 27(4), 433-442.
- Goldstein, H. (2004). International comparisons of student attainment: Some issues arising from the PISA study. *Assessment in Education*, 11, 319-330.
- Goldstein, H., & Healy, M.J.R. (1995). The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, A*, 158, 175-177.
- Goldstein, H., & Spiegelhalter, D. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance (with discussion). *Journal of the Royal Statistical Society, A*, 159, 385-443.
- Goldstein, H., & Thomas, S. (1996). Using examination results as indicators of school and college performance. *Journal of the Royal Statistical Society, A*, 159, 149-163.
- Greaney, V., & Kellaghan, T. (1996). *Monitoring the learning outcomes of education systems*. Washington, DC: The World Bank Institute.
- Hanushek, E.A. (2004). *Some simple analytics of school quality*. NBER Working Paper 10229. Cambridge, MA: National Bureau of Economic Research Inc.
- Hanushek, E.A. (2005a). *Economic outcomes and school quality*. Education Policy Series, Volume 4. Paris: International Institute for Educational Planning and International Academy of Education.
- Hanushek, E.A. (2005b). The economics of school quality. *German Economic Review*, 6(3), 269-286.
- Hanushek, E.A. (2005c). Why quality matters in education. *Finance and Development* (June 2005), 15-19.
- Hanushek, E.A., & Jorgenson, D.W. (Eds.) (1996). *Improving America's Schools: The role of incentives*. Washington, DC: American Educational Research Association.
- Hanushek, E.A., Kain, J.F., O'Brien, D.M., & Rivkin, S.G. (2005). *The market for teacher quality*. NBER Working Paper 11154. Cambridge, MA: National Bureau of Economic Research Inc.
- Hanushek, E.A., & Raymond, M.E. (2004). *Does school accountability lead to improved student performance?* NBER Working Paper 10591. Cambridge, MA: National Bureau of Economic Research Inc.
- Hanushek, E.A., Rivkin, S.G., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica*, 73(2), 417-458.
- Hattie, J.A. (1987). Identifying the salient facets of a model of student learning: A synthesis of meta-analyses. *International Journal of Educational Research*, 11(2), 187-212.
- Hattie, J.A. (2003, October). *Teachers make a difference: What is the research evidence?* Background paper to invited address presented at the 2003 ACER Research Conference, Carlton Crest Hotel, Melbourne, Australia, October 19-21, 2003. Available at:

- <http://www.acer.edu.au/documents/TeachersMakeaDifferenceHattie.doc>.
- Hattie, J.A. (2005). What is the nature of evidence that makes a difference to learning? *Research Conference 2005 Proceedings* (pp. 11-21). Camberwell, VIC: Australian Council for Educational Research. Available at: <http://www.acer.edu.au>.
- Hattie, J., Biggs, J., & Purdie, N. (1996). Effects of learning skills interventions on student learning: A meta-analysis. *Review of Educational Research*, 66(2), 99-136.
- Heckman, J.J. (2000). *Invest in the very young*. Working Paper, Harris Graduate School of Public Policy Studies. Available at: <http://www.HarrisSchool.uchicago.edu>.
- Heckman, J.J. (2005). *Lessons from the technology of skills formation*. National Bureau of Economic Research, Working Paper #1142, February 2005.
- Hill, P.W. (1995). Value added measures of achievement. *IARTV Seminar Series*, No. 44, May, 1995.
- Hill, P.W., & Crévola, C.A. (1999). The role of standards in educational reform for the 21st century. In D.D. Marsh (Ed.), *ASCD Year Book 1999: Preparing our schools for the 21st century* (pp. 117-142). Alexandria, VA: Association for Supervision and Curriculum Development.
- Hill, P.W., & Rowe, K.J. (1996). Multilevel modeling in school effectiveness research (Leading article). *School Effectiveness and School Improvement*, 7(1), 1-34.
- Hill, P.W., & Rowe, K.J. (1998). Modeling student progress in studies of educational effectiveness. *School Effectiveness and School Improvement*, 9(3), 310-333.
- Hoad, K-A., Munro, J., Pearn, C., Rowe, K.S., & Rowe, K.J. (2005). *Working Out What Works (WOWW) Training and Resource Manual: A teacher professional development program designed to support teachers to improve literacy and numeracy outcomes for students with learning difficulties in Years 4, 5 and 6*. Canberra, ACT: Australian Government Department of Education, Science and Training; and Camberwell, VIC: Australian Council for Educational Research.
- Ingvanson, L. (2002). *Development of a National Standards Framework for the teaching profession*. An Issues paper prepared for the MCEETYA Taskforce on Teacher Quality and Educational Leadership. Camberwell, VIC: Australian Council for Educational Research.
- Ingvanson, L.C. (2003). A professional development system fit for a profession. In V. Zbar and T. Mackay (Eds.), *Leading the education debate: Selected papers from a decade of the IARTV Seminar Series* (pp. 391-408). Melbourne, VIC: Incorporated Association of Registered Teachers of Victoria (IARTV).
- Ingvanson, L.C. (Ed.). (2005). *Assessing teachers for professional certification: The National Board for Professional Teaching Standards*. Amsterdam: Elsevier Science.
- Ingvanson, L., Elliot, A., Kleinhenz, E., & McKenzie, P. (2006). *Toward a national approach to the accreditation of pre-service teacher education courses*. A discussion paper prepared for Teaching Australia – the Australian Institute for Teaching and School Leadership. Camberwell, VIC: Australian Council for Educational Research.
- Kellaghan, T., Madaus, G.F., & Airasian, P.W. (1992). *The effects of standardized testing*. Boston: Kluwer-Nijhoff Publishing.
- Lacey, C., & Lawton, D. (1981). *Issues in evaluation and accountability*. London: Methuen.
- Lokan, J., Greenwood, L., & Cresswell, J. (2001). *15-up and counting, reading, writing, reasoning: how literate are Australia's students: The PISA 2000 survey of students' reading, mathematical and scientific skills*. Camberwell, VIC: Australian Council for Educational Research.
- Manno, V.B. (1994). *Outcomes-based education: Miracle, cure or plague?* Hudson Institute Briefing Paper No. 165, June 1994.
- Marks, G.N. (2000). *The measurement of socioeconomic status and social class in the LSAY project*. Longitudinal Surveys of Australian Youth (LSAY), Technical Paper No. 14. Camberwell, VIC: Australian Council for Educational Research.
- Marks, G.N. (2005). Cross-national differences and accounting for social class inequalities in education. *International Sociology*, 20(4), 483-505.
- Marks, G.N. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research*, 48(1), 21-40.
- Marsh, H.W., Rowe, K.J., & Martin, A. (2002). PhD students' evaluations of research supervision: Issues, complexities and challenges in a nationwide Australian experiment in benchmarking universities (Leading article). *Journal of Higher Education*, 73(2), 313-348.

- Masters, G. (1990). Improving the assessment of student outcomes. In J. Hewton (Ed.), *Performance indicators in education: What can they tell us?* (pp. 1-18). Brisbane, QLD: Australian Conference of Directors-General of Education.
- Masters, G.N. (1991). *Assessing achievement in Australian Schools*: A discussion paper commissioned by the Industry Education Forum. Hawthorn, VIC: Australian Council for Educational Research.
- Masters, G.N. (1994). *Setting and measuring performance standards for student achievement*. Paper presented at the conference "Public Investment in School Education: Costs and Outcomes", sponsored by The Schools Council and The Centre for Economic Policy Research, Australian National University, Canberra, March 17, 1994.
- Masters, G.N. (2004). *Continuity and growth: Key considerations in educational improvement and accountability*. Background paper to keynote address Presented at the joint Australian College of Educators and Australian Council for Educational Leaders national conference, Perth, Western
- McGaw, B. (1991). Monitoring education systems. In J. Chapman, L. Angus, G. Burke, & V. Wilkinson (Eds.), *Improving the quality of Australian schools. Australian Education Review No. 33* (pp. 134-139). Hawthorn, VIC: Australian Council for Educational Research.
- Mortimore, P. (1991). School effectiveness research: Which way at the crossroads? *School Effectiveness and School Improvement*, 2 (3), 213-229.
- Muijs, D., & Reynolds, D. (2001). *Effective teaching: Evidence and practice*. London: Paul Chapman Publishing.
- Murphy, P., Greaney, V., Lockheed, M., & Rojas, C. (Eds.) (1996). *National Assessments: Testing the system*. Washington, DC: The World Bank Institute.
- Nisbet, J. (1993). Introduction. In *OECD - Curriculum reform: Assessment in question* (pp. 25-38). Paris: Organisation for Economic Cooperation and Development.
- Nuttall, D.L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, 13(7), 769-776.
- OECD (1983). *Compulsory schooling in a changing world*. Paris: Organisation for Economic Cooperation and Development.
- OECD (1986). *Education and training for manpower development*. Paris: Organisation for Economic Cooperation and Development.
- OECD (1989). *Schools and quality: An international report*. Paris: Organisation for Economic Cooperation and Development.
- OECD (1993). *Curriculum reform: Assessment in question*. Paris: Organisation for Economic Cooperation and Development.
- OECD, (1995). *Indicators of education systems: Measuring the quality of schools*. Paris: Organization for Economic Cooperation and Development.
- OECD (2001). *Teachers for tomorrow's schools: Analysis of the World Education Indicators, 2001 edition*. Paris: Organisation for Economic Cooperation and Development and UNESCO Institute for Statistics.
- OECD (2002). Improving both quality and equity: Insights from PISA 2000. In *Education Policy Analysis 2002* (pp. 35-63). Paris: OECD.
- OECD (2003). *The PISA 2003 assessment framework: Reading, mathematical and scientific literacy*. Paris: Organisation for Economic Cooperation and Development.
- OECD (2005a). *Education at a glance: OECD indicators 2005*. Paris: Organization for Economic Cooperation and Development.
- OECD (2005b). *Teachers matter: Attracting, developing and retaining effective teachers*. Paris: Organization for Economic Cooperation and Development.
- Plomp, T. (1999). *The relevance of IEA type international comparative assessments of educational achievement*. Paper presented at the 40th General Assembly of IEA Oslo, 30 August 1999.
- Rothman, S. (2002). Achievement in literacy and numeracy by Australian 14 year-olds, 1975-1998. *Longitudinal Surveys of Australian Youth, Research Report No. 29*. Camberwell, VIC: Australian Council for Educational Research. Available for download in PDF format at: <http://www.acer.edu.au/research/lsey/reports/LSAY29.pdf>.
- Rowe, K.J. (1989). The commensurability of the general linear model in the context of educational and psychosocial research. *Australian Journal of Education*, 33, 41-52.

- Rowe, K.J. (2000). Assessment, league tables and school effectiveness: Consider the issues and let's get real! *Journal of Educational Enquiry*, 1(1), 72-97.
- Rowe, K.J. (2001a). Educational performance indicators. In M. Forster, G.N. Masters and K.J. Rowe, Measuring learning outcomes: Options and challenges in evaluation and performance monitoring (Revised edition, pp. 2-20). *Strategic Choices for Educational Reform; Module IV – Evaluation and Performance Monitoring*. Washington, DC: The World Bank Institute.
- Rowe, K.J. (2001b, October). *Responsible and irresponsible uses of 'value-added' measures in educational assessment and reporting*. Invited keynote address presented at the Sixth National Roundtable on Assessment and Reporting, Hobart, October 24-26, 2001.
- Rowe, K.J. (2003). *The importance of teacher quality as a key determinant of students' experiences and outcomes of schooling*. Background paper to keynote address presented at the 2003 ACER Research Conference, Carlton Crest Hotel, Melbourne, 19-21 October 2003. Available in PDF format at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J. (2004a). *Analysing and reporting performance indicator data: 'Caress' the data and user beware!* Background paper to invited address presented at the 2004 Public Sector Performance & Reporting Conference (under the auspices of the International Institute for Research – IIR), Sydney, 19-22 April 2004. Available at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J. (2004b). *The importance of teaching: Ensuring better schooling by building teacher capacities that maximize the quality of teaching and learning provision – implications of findings from the international and Australian evidence-based research*. Background paper to invited address presented at the *Making Schools Better* summit conference, Melbourne Business School, the University of Melbourne, 26-27 August 2004. Available in PDF format at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J. (2005a). *Performance indicators of quality schooling: Their nature, construction and use*. Background paper of presentation to senior educational administrators, the Philippines, 21 November, 2005. Camberwell, VIC: Australian Council for Educational Research.
- Rowe, K.J. (2005b). Evidence for the kinds of feedback data that support both student and teacher learning. *Research Conference 2005 Proceedings* (pp. 131-146). Camberwell, VIC: Australian Council for Educational Research [ISBN 0-86431-684-4]. Available for download in PDF format at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J. (Chair) (2005c). *Teaching reading: Report and recommendations*. Report of the Committee for the National Inquiry into the Teaching of Literacy. Canberra, ACT: Australian Government Department of Education, Science and Training. Available at: www.dest.gov.au/nitl/report.htm.
- Rowe, K.J. (2006a). *School performance: Australian State/Territory comparisons of student achievements in national and international studies*. Camberwell, VIC: Australian Council for Educational Research. Available in PDF format at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J. (2006b). *Effective teaching practices for students with and without learning difficulties: Constructivism as a legitimate theory of learning AND of teaching?* Background paper to Keynote address presented at the NSW DET Office of Schools Portfolio Forum, Wilkins Gallery, Sydney, 14 July 2006. Available at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J., & Hill, P.W. (1996). Assessing, recording and reporting students' educational progress: The case for 'Subject Profiles'. *Assessment in Education*, 3(3), 309-352.
- Rowe, K.J., & Hill, P.W. (1998). Modeling educational effectiveness in classrooms: The use of multilevel structural equations to model students' progress. *Educational Research and Evaluation*, 4(4), 307-347.
- Rowe, K.J., & Ingvarson, L. (in press). Conceptualising and evaluating teacher quality: Substantive and methodological issues. *Proceedings of Workshop on Teacher Quality*. Canberra, ACT: Australian National University.
- Rowe, K.J., & Lievesley, D. (2002, April). *Constructing and using educational performance indicators*. Background paper to keynote address and workshops presented at the inaugural Asia-Pacific Educational Research Association (APER) regional conference, ACER, Melbourne, April 16-19, 2002; available at: <http://www.acer.edu.au/research/programs/learningprocess.html>.
- Rowe, K.J., & Rowe, K.S. (1999). Investigating the relationship between students' attentive-inattentive behaviors in the classroom and their literacy progress. *International Journal of Educational Research*, 31(1/2), 1-138 (Whole Issue). Elsevier Science, Pergamon Press.

- Rowe K.J., & Stephanou, A. (2001, October). 'Value-added' educational performance indicators using measurement and multilevel modelling. Invited workshop presented at the Sixth National Roundtable on Assessment and Reporting, Hobart, October 24-26, 2001.
- Rowe, K.J., & Stephanou, A. (2003). *Performance audit of literacy standards in Victorian Government schools, 1996-2002*. A consultancy report to the Victorian Auditor General's Office. Melbourne, VIC: Australian Council for Educational Research.
- Rowe, K.J., Turner, R., & Lane, K. (2002). Performance feedback to schools of students' Year 12 assessments: The *VCE Data Project*. In A.J. Visscher and R. Coe (Eds.), *School improvement through performance feedback* (pp. 163-190). Lisse, The Netherlands: Swetz & Zeitlinger.
- Scheerens, J., & Bosker, R. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.
- StatSoft, Inc. (2005). *STATISTICA data analysis software system* (Version 7.2). Tulsa, AR: StatSoft Incorporated (www.statsoft.com).
- Stephanou, A., & Rowe, K.J. (2002). *WAMSE-English 2001: Multilevel modelling*. A report to the Department of Education, Western Australia, of key measurement and multilevel modelling findings related to the *Reading* and *Viewing* achievements of students in Years 3, 7 and 10. Camberwell, VIC: Australian Council for Educational Research.
- Summit of the America's – Line 2: Educational Assessment, Brasilia*, March 12-14, 2002. INEP.
- Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years (Leading article). *School Effectiveness and School Improvement*, 8, 169-197.
- Thomson, S., Cresswell, J., & De Bortoli, L. (2004). *Facing the future: A focus on mathematical literacy among Australian 15-year-old students in PISA 2003*. Camberwell, VIC: Australian Council for Educational Research.
- Thomson, S., Rowe, K.J., Underwood, C., & Peck, R. (2005). *Numeracy in the early years: Project Good Start*. Final report to the Australian Government Department of Education, Science and Training. Camberwell, VIC: Australian Council for Educational Research. Available at: http://www.dest.gov.au/sectors/school_education/publications_resources/profiles/goodstart.htm#publication: and at: <http://www.acer.edu.au/research/projects/goodstart/documents/GoodstartFinalReport.pdf>.
- Tucker, M.S., & Codding, J.B. (1998). *Standards for our schools: How to set them, measure them and reach them*. San Francisco, CA: Jossey-Bass.
- Visscher, A.J., & Coe, R. (Eds.) (2002). *School improvement through performance feedback*. Lisse, The Netherlands: Swetz & Zeitlinger.
- Visscher, A., Karsten, S., de Jong, T., & Bosker, R. (2000). Evidence on the intended and unintended effects of publishing school performance indicators. *Evaluation and Research in Education*, 14, 254-267.
- Watson, L. (1996). Public accountability or fiscal control? Benchmarks of performance in Australian schooling. *Australian Journal of Education*, 40, 104-123.
- Westwood, P.S. (2004). *Learning and learning difficulties: A handbook for teachers*. Camberwell, VIC: Australian Council for Educational Research.
- Westwood, P.S. (2006). *Teaching and learning difficulties: Cross-curricular perspectives*. Camberwell, VIC: Australian Council for Educational Research.
- Willms, J.D. (2000). Monitoring school performance for standards-based reform. *Evaluation and Research in Education*, 14, 237-253.