NSW GOVERNMENT Education & Communities

TECHNICAL PAPER / JULY 2014

Value added models for NSW government schools

Centre for Education Statistics and Evaluation

Publication and Contact Details

Centre for Education Statistics and Evaluation 2014,
Value added models for NSW government schools,
Report prepared by L Lu & K Rickard. This report is available at:
http://www.cese.nsw.gov.au/publications/reports-and-presentations

Authors:

Karen Rickard

Statistician | Statistics Unit
Centre for Education Statistics and Evaluation
Strategic Information and Reporting
Office of Education
NSW Department of Education and Communities
T 02 9561 1234
F 02 9561 8055

Dr. Lucy Lu

Leader | Statistics Unit
Centre for Education Statistics and Evaluation
Strategic Information and Reporting
Office of Education
NSW Department of Education and Communities
T 02 9561 8691
F 02 9561 8055

NSW GOVERNMENT Education & Communities

Contents

Executive Summary

Identifying high performing schools is an important step in developing the evidence base about "what works" to improve educational outcomes for students. However, such a task is not straightforward. Absolute performance measures (e.g., average test scores of a school) are commonly used by the media to develop league tables of schools. Such measures are generally more reflective of the characteristics of students attending a school, rather than the contribution the school has made to its students' learning. The Centre for Education Statistics and Evaluation (CESE) has developed a suite of value added (VA) measures that are intended to be fair and robust indicators of the contribution schools make to their students' development. The value added measures take into account those contextual factors (both school-and student-related) that impact on students' learning and that are conceived to be largely beyond the control of schools. They help to identify schools that make a larger than average contribution to students' learning, as the basis of further investigation of "what works" to provide sound evidence for educators and policy makers to continually improve teaching and learning.

Based on available state-wide student assessment data, VA measures have been developed using the following matched student test results: Year 3 to Year 5 (NAPLAN), Year 5 to Year 7 (NAPLAN), Year 7 to Year 9 (NAPLAN), and Year 9 to Year 12 (NAPLAN to HSC). In addition, an exploratory Kindergarten to Year 3 measure (Best Start to NAPLAN) is also being trialled.

The key features of the proposed DEC VA models include:

- Explicitly accounting for the available school and student contextual factors that have been shown to have a persistent and significant impact on students' learning outcomes
- Utilising a multilevel modelling approach that takes account of the nesting of students within schools and hence provides more reliable and accurate school effect estimates
- Pooling of data across two measurement periods to reduce random errors, so that estimates are more likely to reflect any persistent differences in school performance
- Reducing the volatility of VA estimates, for small schools especially, by applying a statistical technique that adjusts the estimates in proportion to their reliability.

Future work planned to further enhance the validity and reliability of the VA measures includes:

- Estimating bias arising from movements of students across schools and across sectors within a measurement period
- Investigation of the suitability of teacher assessments at entry to school (e.g., Best Start program data) as a reliable baseline indicator for the VA K-3 measures
- Ongoing work to identify other, currently unmeasured, contextual factors (such as student mobility and student disability).

The suite of VA measures that have been developed can help schools to evaluate their performance and to identify and implement improvement strategies. However, care needs to be taken when interpreting and using VA estimates. It is recommended that VA estimates are always reported with confidence intervals, and along with estimates for previous years where possible. Guidelines to aid the interpretation of VA estimates should also be developed and included in the reporting package so that schools can make the best use of the VA information.

1 Introduction

The Centre for Education Statistics and Evaluation (CESE) within the NSW Department of Education and Communities (DEC) is charged with developing the evidence base about "what works" to improve educational outcomes for students. One route to discovering what works is to identify high performing schools and to analyse what it is they do. However, identifying high performing schools is not a straightforward task. Absolute performance measures, such as average test scores or the percentage of students in the highest performance bands, are commonly used by the media to develop league tables of schools. Such measures, however, are generally more reflective of factors such as school academic selectivity and students' socio-economic background, rather than the contribution the school has made to its students' learning. Instead, fairer measures of school performance are required that level the playing field, by taking into account those contextual factors (both school- and student-related) that impact on students' learning outcomes and that are beyond the control of schools.

In addition to identifying schools that contribute the most to students' learning development, fairer measures of school performance could have much broader application. They would enable all schools to better evaluate their own performance and hence contribute to school improvement strategies and practices, as well as potentially contributing to school accountability and/or accreditation mechanisms.

The fairer measures of school performance that are proposed for NSW government schools are referred to as value added (VA) measures. A VA measure estimates the effect a school has in improving the learning progress of a particular cohort of students over a particular time period (e.g., the cohort of students who progressed from Year 3 in 2010 to Year 5 in 2012) relative to other schools. Compared to absolute performance measures or raw growth data[1], the VA measures proposed explicitly adjust for contextual factors that have been shown to have a persistent and significant impact on student learning outcomes, such as differences in student background (e.g., socio-economic status (SES)) and in school environments such as the concentration of low SES students or the selection of high achieving students in a school due to admission policies set by the system. By removing the external and contextual influences, VA measures seek to provide a more accurate (and fairer) indication of a school's contribution to the development of their students over a period of time.

The development of the VA measures has proceeded through a number of (iterative) steps which included a review of the literature on school performance models, developing/refining models, testing model parameters, reviewing and sense checking results with school educators and testing the stability of VA estimates using data from available years. Key considerations underpinning our modelling process include validity (can we draw valid inferences from the measures related to their intended purpose/s), reliability (how much confidence we have with these measures), stability (how volatile they are from year to year) and consistency (can we draw consistent interpretations from the VA measures developed for different grade cohorts in a school or over time). These considerations are further detailed in the subsequent sections as the rationale for each of our modelling choices is explained.

Our value added modelling has also benefitted from the expertise of international and Australian academic experts on the independent Advisory Council for the Centre for Education Statistics and Evaluation, who provided invaluable advice on methodological issues when consulted.

This paper aims to: provide an overview of the VA models we developed for DEC schools; present the rationale behind the model specifications; and discuss the appropriate interpretation, proposed reporting, and use of VA measures.

1 Student growth data from NAPLAN testing is provided by both ACARA via the My School website, and by DEC to schools via SMART (an analytical package schools use for evaluation, planning and improvement purposes). The growth data presented allows comparisons with the average growth of schools with similar levels of student socio-educational advantage or with the average growth of students with similar starting points, but does not account for the impact of SES and other student and school level contextual factors on students' learning gains simultaneously. This is critical for fair assessment of the contribution a school has made to student learning. See Section 3.1 for further discussion.

2 Concept of value added measures

Over the past four decades, a voluminous body of research into school effectiveness has been conducted seeking to identify the level of influence schools and teachers exert on students' achievements. Results from these studies (e.g., Ballou, Sanders & Wright, 2004; Correnti & Miller, 2002; Cuttance, 2000; Mujis & Reynolds, 2001; Rowan, Correnti & Miller, 2002; Rowe, 2004) have repeatedly demonstrated that a large portion of the variation in the average performance of students attending different schools is attributable to the differences in the nature of the students attending the schools, rather than differential school effectiveness. Characteristics of students such as prior achievement levels, language proficiency of students learning English as an Additional Language or Dialect (EAL/D), and socio-economic status have all been shown to be important factors impacting on students' achievements and their rates of learning progression. Since students are not randomly assigned to schools, fair assessment of school effectiveness must take account of these 'non-school' (contextual) factors.

In addition to the above mentioned student characteristics, school level contextual factors such as the social composition of the student body in the school, the location of the school (whether in a metropolitan or remote area), and type of school (a single sex or a co-educational school; a non-selective or a selective school which streams students for admission under the DEC enrolment policy), can also play a part in the students' learning outcomes. These external and contextual factors are considered beyond the control of school principals and staff, therefore their influences on students' scores (if significant) should also be adjusted for in a school performance measure so that every school can be assessed on a fair and equal basis.

The development of value added models stems from this need to consider external influences when assessing school effectiveness. While there is not a single unifying definition of the term 'value added', in education, it is most commonly used to "describe the additional value schools bring to the learning outcomes of their students" (DEECD, 2007, p. 3). The essence of the term is well captured by Hill (1995) when he described the VA measures as those that attempt to "indicate the educational value that the school adds over and above that which could be predicted given the backgrounds and prior attainments of the students within the school" (p. 6).

Typically, value added measures for schools (also interchangeably referred to as 'school effects' in this paper) are estimated for a particular student cohort over a particular period of time, by first calculating the difference between the actual student outcome of interest and the outcome predicted using a statistical method based on the student prior achievement level and other student background and school contextual factors. This difference is then statistically averaged across all students from the same target cohort attending a school to derive a summary school level measure. For example, a Year 7 to 9 value added measure for a high school for the time period 2010 to 2012 seeks to quantify the average gain the target cohort of students achieved as they progressed from Year 7 in 2010 to Year 9 two years later, relative to the gain these same students would have achieved had they enrolled in an 'average' school serving students of similar background.

Two key features of the school VA measures are therefore important to stress. First, these measures are not about student achievement level, but about shedding light on the effectiveness of schools. Secondly, the measures are inherently of a relative nature. They are not absolute measures of schools' contributions to students learning; rather they compare the impact of a school relative to other schools. As pointed out by Johnson et al. (2012, p. 1), the key question addressed by these measures is "To what extent does the actual level of student performance exceed (or fall short of) the level that is predicted for students with similar achievement histories and background characteristics if taught by the 'average' teacher or school?"

3 Review of different types of VA models

VA models were first implemented as part of the operation of a school system in Tennessee, USA in the early 1990's, initially as a test based accountability mechanism (CGP, 2004). Since then, VA models have gained an increasingly wider acceptance amid growing demands for accountability of schools and teachers (Deming, 2014). The international and Australian educational systems that have adopted a VA approach to routinely provide estimated school/teacher effects to schools and/or the public include the United Kingdom, Hong Kong, more than 30 USA states/districts (e.g., Tennessee, Dallas, North Carolina, South Carolina, Pennsylvania, Arkansas, Minnesota, Massachusetts, Ohio, District of Columbia) and locally in Australia, the Victorian Department of Education and Early Childhood Development and NSW Catholic education system (Blank, 2010; Hershberg, Adams, Simon & Lea-Kruger, 2004; Leckie, 2013; Hong Kong Education Bureau, 2012). The VA information is used by these systems for a range of purposes, from relatively low stakes use such as informing school and teacher improvement, school self-evaluation, and monitoring policy initiatives to high stakes use (e.g., as an accountability mechanism).

The VA models implemented so far in Australia and internationally vary in the statistical techniques used to calculate school effects as well as in the model compositions. This is not surprising, given that they were developed at different times and for potentially different purposes. A multitude of factors could have influenced the modelling choices: 1) the policy and political environment which could impact on the inclusion and exclusion of certain controlling factors in the VA models; 2) the statistical modelling techniques available at the time; 3) data availability; and 4) different views on the definition of school effectiveness and the underlying assumptions about students' growth.

With varying complexities, the main types of VA models that have been adopted by international and Australian educational systems include:

3.1 Gain score model

The simplest form of this model is one that uses the average of the gain scores (differences in the 'pre'-test scores and 'post'-test scores) across all students in a school. This type of the gain score model is used by ACARA (Australian Curriculum, Assessment and Reporting Authority) in its reporting of 'student gain' on My School.

A more complex form is the Student Growth Percentile model implemented in the US state of Massachusetts (MCAS, 2011). This model calculates the growth percentile for each student, by comparing the growth of a student in a school to the average growth of all students in the district/system with similar scores in previous tests. The growth measure is then aggregated across all students in the school as an indicator of the value a school adds to its students' learning.

A key assumption of these models is that the impact of the external and contextual factors (e.g., student SES) on students' attainment has been fully accounted for in their prior achievement scores and that these factors do not have any residual impact on student growth. This assumption is not always met, as demonstrated in the following graph using the expected growth data from the DEC SMART.

In Figure 1, each dot represents a DEC primary school. The horizontal (x) axis represents a school's Family Occupation and Education Index (FOEI) [2] value, which is a proxy for the average socio-economic status of the students at the school. The greater the FOEI value, the lower the school's SES. The vertical (y) axis is the proportion of Year 5 students in a school achieving at or above expected growth[3].

2 FOEI is derived from the parental education and occupation information provided by the parents on student enrolment forms. Further information about FOEI can be found at: http://www.cese.nsw.gov.au/publications/learning-curve/item/38-learning-curve-5.

3 A student is defined as achieving at or above expected growth if the student's growth is either within 0.2 standard deviations of the mean growth (i.e. "at" expected growth) or greater than 0.2 standard deviations above the mean growth (i.e. "above" expected growth) of all students in the system who had the same starting point.

Figure 1:

Relationship between the percentage of Year 5 students achieving at or above expected growth and school SES

Source: NSW government schools growth data (2012) from SMART. School SES is represented by the latest (2013) Family Occupation and Education Index (FOEI) values for government schools.

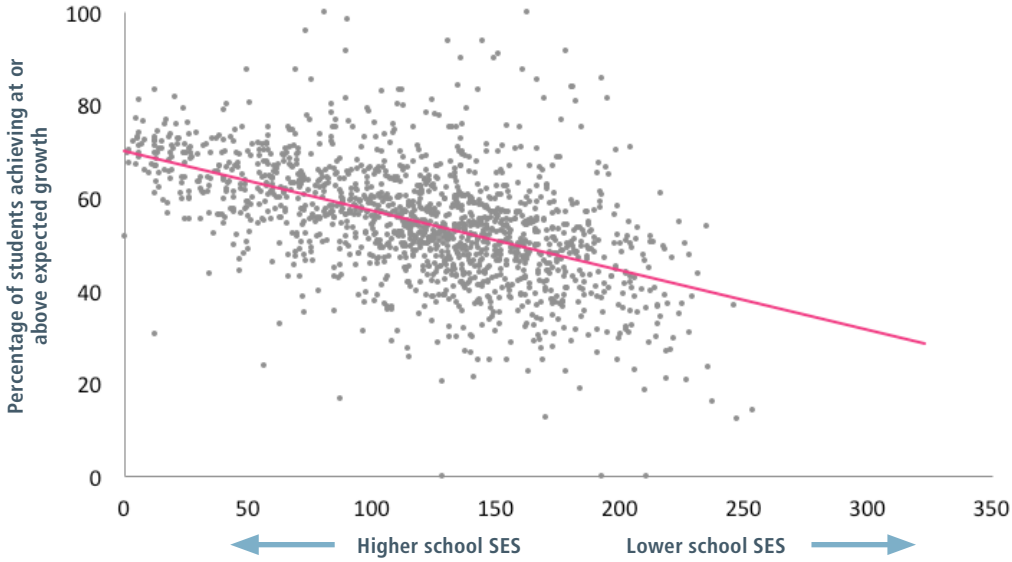


Figure 1 shows that higher SES schools tend to have higher growth, relative to their students' starting points[4]. This indicates that the effect school SES has on students' growth is cumulative, and is not fully accounted for in students' prior scores. Note that many of the outliers in the graph are small schools, whose growth data are more influenced by the idiosyncratic performance of a few students than large schools.

Our analysis of NSW government school data also shows that gain score measures can be fairly unreliable and volatile for small schools, due to the relatively high level of random error to which small schools' gain score measures are subject. This results in the over-representation of small schools at the extreme ends of the gain distribution. For example, 90 per cent of schools with average gain scores[5] more than 2 standard deviations above or below the average are schools with fewer than 10 matched student records. Furthermore, one in five small schools[6] switch from significantly above average gain in one year to significantly below average gain in the next year, or vice versa. The equivalent rate of switching for schools with more than 50 matched student records is less than 3 per cent. On the basis that there is no strong reason to believe the rate of performance improvement or deterioration for a small school is any different than that for a large school, such volatility in the school performance ranking using raw growth data shows the extent to which the reliability of school ranking based on raw growth data can be adversely influenced by the small number of the underlying observations.

3.2 Covariate adjustment model (single level regression model either at the student or school level)

While gain score is the outcome of interest in the previous model, the covariate adjustment model uses the current test score as the outcome variable, and attempts to explain the variation in this variable by using a simple statistical regression method to account for prior achievement of students attending a school and other student and/or school level contextual factors. An example of such a model (see DEECD, 2011) is one that uses a regression technique to predict mean school performance based on the average characteristics of the students in the school (e.g., school average prior achievement and average background factors such as percentage of low SES students). The difference between the actual school mean performance and the predicted mean performance is then interpreted as an indicator of school effectiveness.

4 A similar pattern also exists for secondary schools.

5 This analysis uses school average gain scores, which are derived by first calculating the difference between the average test score a student achieved in Year 3 2011 over reading and numeracy and the equivalent average test score the same student achieved two years later in Year 5, and then aggregating this difference over all students attending a school.

6 For this analysis, a small school is defined as one with fewer than 10 matched student records (i.e. the number of Year 3 students in 2011 who were able to be matched to the Year 5 NAPLAN cohort in the same school in 2013).

Another form of this model is one that predicts students' current scores based on a set of student level characteristics, and then uses the average difference between the predicted and actual scores aggregated across all students attending a school as a measure of school performance (OECD, 2008).

The main drawback of this model is that it does not take into account the hierarchical nature of educational data (e.g., students are nested within schools). The adverse statistical consequences of ignoring the nesting structure of the data has been studied and discussed extensively in the literature (e.g., Aitkin & Longford, 1986; Rasbash et al., 2005; Raudenbush & Bryk, 2002). The main conclusion is that such a modelling choice can produce mis-estimated standard errors, potentially misleading school effect estimates, and for schools with small sample sizes, unstable estimates.

3.3 Multilevel models

Although a more complex approach, the multilevel modelling approach is statistically superior to the previous two methods as it takes into account the clustering nature of education data and allows for many levels of effects to be simultaneously modelled: 1) intra student – performance across all tests in the current or previous years; 2) student (e.g., characteristics); 3) teacher; and 4) school. These models are therefore better able to isolate the effect we want to measure from all other effects and produce value added measures that are less subject to statistical bias (Rowley, 2006).

The main types of multilevel VA models that have been implemented in other educational systems seem to fall into two categories:

a) Models that rely solely on the previous and current test scores, such as the Educational Value added Assessment System (EVAAS) implemented in Tennessee, USA (Sanders, 2000). The EVAAS model compares the actual progress of each student in a school to the expected progress for that student, which is obtained through complex analysis of the variance and covariance structure in students' test scores across different subjects and between different grades (Ballou et al., 2004; Rowley, 2006). Apart from students' test scores, the model does not further adjust for any non-educational factors. This is based on the assumption (Sanders, 2000) that by accounting for prior achievement, the model has fully controlled for the influence of other background factors such as socio-economic factors on students' subsequent achievement. DEC analysis using the expected growth data (see p. 7) supports other researchers' argument that there is insufficient evidence backing such a claim (Griffin, Woods & Nguyen, 2005; OECD, 2008). The model also has significant resource implications as the extraction of variance-covariance structure depends on an extensive testing regime that involves testing students in multiple subjects every year.

b) Models that compare the progress students in a school make, relative to all other students with similar prior achievement, while controlling for student and school contextual factors. These models are generally referred to as multilevel contextual value added models and have been implemented in the United Kingdom (Leckie, 2013), Hong Kong (HK Education Bureau, 2012) and a number of states/districts in the USA (e.g., Isenberg & Hock, 2011; Johnson et al., 2012; Webster, 2005[7]). Details of this type of model are explained in greater depth in the next section.

7 Strictly speaking, the Dallas model, presented in Webster (2005), uses a combination of different types of models mentioned in this section. It is a two stage model, whereby in the first stage, student test scores (current and prior scores) are adjusted using a single level regression model controlling for a set of student level characteristics. In the second stage, the adjusted current score (from the first stage) was regressed on the adjusted prior scores using a multilevel modelling technique that takes into account the nesting of students within schools, as well as controls for a set of school level contextual factors (OECD, 2008).

4 Proposed VA models for NSW DEC

After having assessed the advantages and disadvantages of different types of VA models and having conducted a range of modelling exercises using NSW government school data, we propose using a suite of multilevel (two-level) contextual value added models, each developed for a specific target cohort of students in a school over time, to generate a profile of VA measures for every mainstream school.

The range of VA measures possible for NSW government schools is dependent on the availability of consistent system-wide student assessment data. Unlike many of the states in the USA where annual test scores for students are often available, NSW government schools do not have system-wide standardised tests of students' achievement on an annual basis. Rather students are tested at two-year intervals, when they are in Years 3, 5, 7 and 9, through the National Assessment Program – Literacy and Numeracy (NAPLAN)[8]; and later when they are in Year 12 through the NSW Higher School Certificate (HSC) examinations in various subjects. In addition, for NSW government schools, all Kindergarten students are assessed by classroom teachers at the beginning of the year over multiple aspects of literacy and numeracy to enable teachers to gain a good understanding of the learning needs of every student at their entry to school. With primary schooling in NSW ranging from Kindergarten to Year 6, and secondary schooling covering Years 7 to 12, it would appear then at least four VA measures could be developed as indicators of the value schools add to students' learning at different learning stages: for primary schools, from Kindergarten to Year 3 (hereafter referred to as VA K-3), and from Year 3 to Year 5 (VA 3-5); and for secondary schools, from Year 7 to Year 9 (VA 7-9) and from Year 9 to Year 12 (VA 9-12).

Further, as DEC can track the movement of students from Year 6 to Year 7 if they stayed in the government system, it is also possible to construct a VA 5-7 measure, based on a primary school's Year 5 students' test results and the corresponding results of this cohort two years later from the Year 7 NAPLAN tests. The VA 5-7 measure could be construed as the value added by the primary school to its students over their last two years of primary education, since over 90 per cent of the period between the Year 5 NAPLAN testing time and Year 7 NAPLAN testing time occurred in primary schools.

In all the proposed models, the dependent variables are students' later performance scores (or post-test scores), not the gain scores (or differences in the test scores between two test times). Models using the post-test scores as outcome variables and pre-test scores as a student-level covariate in a multilevel model are sometimes referred to as 'quasi-gain models' (Schochet & Chiang, 2010, p. 3). This modelling choice was necessary for Year 9-12 VA models, as Year 12 HSC results are not equated to the same scale as the Year 9 NAPLAN scores, therefore calculation of direct gain scores is not possible. For the VA models that rely on NAPLAN scores only (i.e., VA 3-5, 5-7 and 7-9), given that the NAPLAN scores are vertically equated across the test grades, value added measures produced from multilevel models using gain scores as dependent variables are typically very similar to those based on the 'quasi-gain' models (Harris & Sass, 2006, as cited in Schochet & Chiang, 2010, p. 3). Therefore in order to keep consistent the interpretation of results from the VA models, students' later test scores are used as dependent variables, and prior scores as covariate variables, in all the proposed VA models.

The specific NAPLAN test scores used in the models are (standardised) test scores aggregated over reading and numeracy. The test scores used for Year 12 students in the main VA 9-12 models[9] are the (standardised) aggregated results across each student's best 10 HSC units. See section 4.9 for more details about the standardisation process and the rationale for using composite scores as the prior and later ability scores. With regard to the VA 9-12 models, it is acknowledged that, for some schools, the proportion of matched students from Year 9 to Year 12 could be small, due to reasons such as students moving schools during the measurement period or leaving the government education system before completing Year 12. This could introduce bias to the VA estimates since students leaving

8 NAPLAN is conducted early in a school year (in May) and tests skills in literacy and numeracy that are developed over time through the school curriculum. For further information on NAPLAN, refer to the website: http://www.nap.edu.au/naplan/naplan.html.

9 Three VA 9-12 measures are developed as a result of the DEC value added modelling exercise. The main measure VA 9-12(TES) attempts to estimate a school's contribution to its senior secondary students' learning by using weighted results on the best 10 HSC units each student attempted as the outcome variable. Two other supplementary measures are also developed – VA 9-12(English) and VA9-12(Maths), which estimate a school's contribution to its senior students' learning in English and Maths subjects respectively.

the school before Year 12 could have different learning trajectories than those who stayed. The development of additional non-test based VA measures (such as using the Year 12 completion rate as the dependent variable), as discussed in Section 6, could provide a more balanced and holistic picture of a school's value added to its students' educational outcomes.

It is also important to note that all the proposed VA models are developed for mainstream schools, the educational excellence of which might be gauged by their students' learning gains measured from national or state literacy and numeracy tests. These models are not intended for Schools for Specific Purposes, which enrol students with disabilities and special needs. Different measures need to be developed to evaluate the educational success of these schools.

The essence of the VA models we are proposing is the use of the multilevel modelling method to estimate the school's relative contribution to students' learning over a time period, having adjusted for initial intake differences (e.g., student starting ability levels) and other student and school level contextual variables.

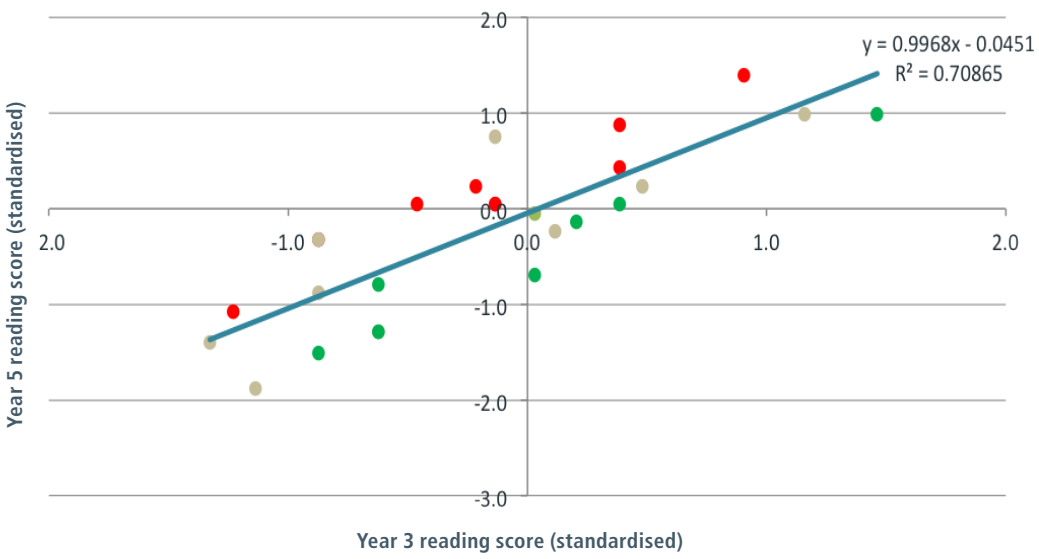
The next section explains the details and the benefits of this modelling technique to the estimation of school effects.

4.1 Introduction to a multilevel model

To explain the multilevel model, we begin by considering a simple (single level) regression model, which examines the relationship in the population between Year 5 students' reading results, say in 2013, and their corresponding results when they were in Year 3, 2011. Figure 2 provides a scatterplot of the relationship between the two variables, based on a sample of matched student records from DEC schools. Red dots indicate students who are attending School A, green dots representing those attending School B and grey dots indicate students attending other schools.

Figure 2:

Scatterplot of students' reading scores in Year 3 (2011) and their matched scores in Year 5 (2013), based on a sample of matched records



We observe that prior achievement is a significant predictor of reading scores two years later. The regression line shown in Figure 2 represents the Year 5 reading scores predicted solely on the basis of students' prior Year 3 reading scores. Mathematically, the regression model in Figure 2 can be written as:

$$y_{i,t} = \beta_0 + \beta_1 * y_{i,t-1} + e_i \qquad [1]$$

Where $y_{i,t}$ is the reading test score of the ith student at time t (e.g., in Year 5, 2013), $y_{i,t-1}$ is the reading test score of the ith student at a previous time $t - 1$ (e.g., in Year 3, 2011). For ease of interpretation, both $y_{i,t}$ and $y_{i,t-1}$ have been standardised to have a distribution with a mean of 0 and standard deviation of 1.

β_0 is called the 'intercept' which is the expected result when the standardised prior test score is zero (i.e., when the student has an average level of prior achievement). e_i is referred to as the 'residual', which is the difference between a student's actual performance and the predicted performance based on the student's prior achievement level. This is the variation in the student test score that cannot be explained by the differences in the students' starting ability. However, when taking a closer look, it is observed that students attending School A tend to perform above the predicted scores (i.e., points lie above the regression line) whereas those attending School B perform below the predicted scores (i.e., points lie below the regression line). This suggests that some of the unexplained variance in students' scores may be further explained by the schools they are attending.

To estimate the school effects, a traditional method is to fit an ANCOVA model, which compares the adjusted mean performance for schools, after holding constant the differences in students' prior achievement across schools.

Assuming there are n schools in the population, and let the n^{th} school be the reference school, then the ANCOVA model can be written as:

$$y_{i,j,t} = \beta_0 + \beta_j + C_1 * y_{i,j,t-1} + e_{ij} \qquad j = 1, \ldots, n\text{-}1 \qquad [2]$$

where $y_{i,j,t}$ is the test score of the i^{th} student in j^{th} school at time t,

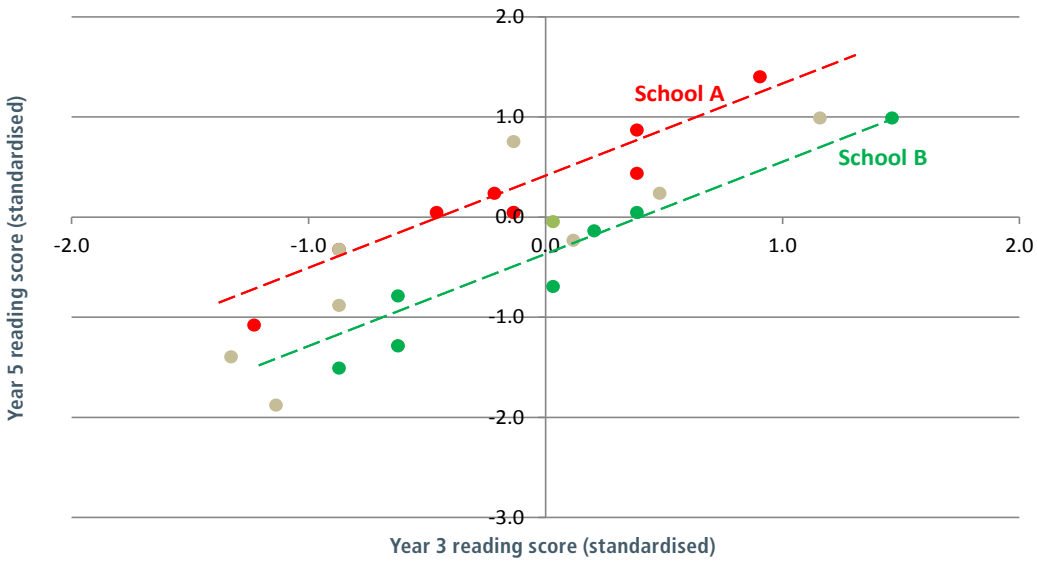
β_j is the coefficient associated with school j (Note: as n is the reference school $\beta_n = 0$)

C_1 is the effect of student's prior achievement on his/her later test score, assumed to be constant across schools

e_{ij} is the residual assumed to be normally distributed with a mean of 0 and variance of σ^2_e.

Model 2 can be visually illustrated in Figure 3, whereby essentially each school now has its own regression line with the same slope but varying intercept $\beta_0 + \beta_j$. The difference in any pair of intercepts between two schools is the difference in the adjusted mean performance between the two schools, after having controlled for the differences in students' starting ability. Assuming students are randomly assigned to schools, β_j can then be considered as the additional value added by a particular school to its students' outcomes, which is our investigation of interest.

Figure 3:
A simple illustration of school effects



The above ANCOVA model can be further extended to include additional student level background variables, if required, to reduce the bias in the school effect estimates introduced by the non-random nature of the assignment of students to schools.

However, this traditional method of estimating school effects has a number of limitations:

a) For small schools (with few students), the estimated effect sizes can be highly unreliable.

b) If there are a large number of schools, which is the case for the NSW government sector[10], the model will need to estimate a large number of parameters (β_j) which can cause statistical issues.

c) The model cannot separate the effect of other school level factors (e.g., school socio-economic status) from the school effects, as they are confounded with the group (school) effects. As students are rarely randomly assigned to schools, any failure to represent these selection artefacts through the model weakens the validity of the school effect estimates.

d) The model does not allow for more complex relationships to be taken into account when estimating school effects. For example, the impact of the prior achievement on students' subsequent scores (C_1 in equation 2) might be significantly different across schools. Lacking the ability to incorporate this variability in the model may impact on the accuracy of the school effect estimates produced by the model.

The multilevel modelling method can deal with all these problems.

4.2 Multilevel model equation

A key difference between a multilevel model and the models discussed above lies in the way school effects are treated. In a multilevel model, schools are considered as a sample drawn from a wider population of schools, and the school effects (represented by μ_{0j} in equations 4 and 5) are normally distributed with a mean of zero and variance $\sigma^2_{\mu 0}$ (Rasbash et al., 2005, p. 30)[11]. The population is considered to have a hierarchical structure with level 1 units (e.g., students) clustered within level 2 units (e.g., schools) (Rasbash et al., 2005).

In an educational context, the simplest multilevel model with no explanatory variables (often referred to as the 'null model') can be written as:

$$\text{Level 1 (student)} \qquad y_{i,j} = \beta_{0j} + e_{ij} \qquad [3]$$

$$\text{Level 2 (school)} \qquad \beta_{0j} = \beta_0 + \mu_{0j} \qquad [4]$$

This model at level 1 predicts the outcome of student i in a school j ($y_{i,j}$) with just one level 2 parameter, the intercept β_{0j}, which is the mean outcome for school j. At level 2, β_{0j} is also construed as varying about the overall population mean of β_0, with μ_{0j} representing each school's difference from this mean.

Substituting [4] into [3], the null model can also be written as:

$$y_{i,j} = \beta_{0j} + \mu_{0j} + e_{ij} \qquad [5]$$

It is easier to see from equation 5 that the residual (unexplained part) in the student scores is now partitioned into two components: (a) the student-level residual (e_{ij}, with variance σ^2_e) which is each student's departure from the predicted outcome, and (b) the school-level residual (μ_{0j}, with corresponding variance $\sigma^2_{\mu 0}$) which is the difference between the school mean and the overall population mean. The variances of these two error terms are known as the within-school variance (σ^2_e) and the between-school variance ($\sigma^2_{\mu 0}$).

The above model can be expanded to include explanatory variables both at the student level and school level. For example, assuming that student i in school j's performance is predicted by one student-level factor X_{1ij} (e.g., student SES) and one school level factor W_{1j} (e.g., remoteness of school), and assuming also that the impact of these factors on the outcome variable is invariant across schools, we can write the multilevel model as:

10 There are over 1600 primary schools and 580 secondary schools in the NSW government system.

11 Since the group effects from these models are considered to be randomly distributed, these models are also referred to as 'random effects models' in the literature. In contrast, models such as the ANCOVA models previously discussed are referred to as 'fixed effects models'

Level 1 (student) $$y_{i,j} = \beta_{0j} + \beta_1 * X_{1ij} + e_{ij}$$ [6]

Level 2 (school) $$\beta_{0j} = \beta_0 + \beta_2 * W_{1j} + \mu_{0j}$$ [7]

Or for simplicity, we can combine the two equations into one:

$$y_{i,j} = \beta_0 + \beta_1 * X_{1ij} + \beta_2 * W_{1j} + \mu_{0j} + e_{ij}$$ [8]

This model is also commonly referred to as the 'random intercept' model since there is only one random variable at the school level, that is the intercept β_{0j} in equation 7.

The above model can be expanded further to investigate more complex problems. For example, we might hypothesise that the impact of the explanatory variable X_{1ij} on the outcome is different across schools. To test this hypothesis, we can write the new equation as:

$$y_{i,j} = \beta_0 + \beta_{1j} * X_{1ij} + \beta_2 * W_{1j} + \mu_{0j} + e_{ij}$$ [9]

where $$\beta_{1j} = \beta_1 + \mu_{1j}$$ [10]

In model 9, β_{1j} (the slope of the regression line for school j for the student-level factor X_{1ij}) is assumed to be a random variable, with μ_{1j} representing the deviation for school j from the average slope in the population (β_1). The significance (and the size) of the variability in the slopes allows us to examine important research questions such as whether some schools might be more equitable than others and whether some schools are differentially effective for different groups of students (see further details in Section 4.7). Since slopes are modelled as randomly varying about their overall means in the population, equation 9 is often referred to as a 'random slope' model.

Compared to the traditional methods, multilevel modelling has a number of statistical advantages which help produce more reliable and accurate school effect estimates. First, while all of the parameters in equation 9 can be estimated through a multilevel model, this is not possible through a classical ANCOVA model (Raudenbush & Bryk, 2002). For example, a classical ANCOVA assumes the homogeneity of regression coefficients (i.e., slopes of student-level predictors are constant across schools), whereas such an assumption can be explicitly tested in a multilevel model. Secondly, the number of parameters that need to be estimated is significantly reduced in a multilevel model than in a fixed effects ANCOVA model, when the number of level 2 units (e.g., schools) is large. Thirdly, when compared to models that ignore the clustering of students in schools, multilevel models produce efficient parameter estimates and appropriate standard errors by properly representing sources of variation in nested designs (Raudenbush & Bryk, 2002). Finally, as multilevel models treat schools as samples from wider populations of schools, they allow us to make inferences beyond the samples. For example, equation 9 enables us to infer the range of the slopes for the explanatory variable X_{1ij} we might expect to see in the whole population of schools.

Having tested both types of models (i.e., the random intercept and random slope models), we propose to use a random intercept model for our purposes of estimating school level effects. Before we expand on the rationale for this choice and produce the comparisons of the results from fitting both types of models in Section 4.7, we first explain the controlling factors (i.e., contextual factors) proposed to be included in our VA models.

4.3 Contextual factors included in the DEC VA models

Selecting the appropriate contextual factors for inclusion in the models is an important step in developing a VA model, for several reasons.

From a statistical point of view, omitting key student background factors that are strong predictors of students' outcomes from the VA models increases unexplained level-1 error variance, which in turn leads to a reduction in the precision of any estimates of school effects and the power of hypothesis tests (Raudenbush & Bryk, 2002). However,

more importantly from a validity perspective, as students are not randomly allocated to schools, failing to adjust for background factors such as differences in the prior achievement and students' socio-economic status can lead to bias in the estimated school effects (Rothstein, 2010). In this regard, a few studies on teacher effects have demonstrated that adjusting for prior test scores and other student characteristics is sufficient to account for non-random assignment of students across teachers within a school (Chetty, Friedman & Rockoff, 2013; Kane & Staiger, 2008; Kane et al., 2013). Similarly, a recent study (Deming, 2014) comparing school effect estimates from an experimental design (i.e., lottery-based random assignment of students to charter schools) to estimates from VA models adjusting for prior test scores and student background factors also found the two sets of estimates to be similar. This suggests that, when important factors are controlled for in a school VA model, assignment of students may be considered at least functionally random, which is essential if causal inferences from school effects are to be made.

Based on reviews of relevant literature and previous DEC analyses, we tested a number of 'non-school' factors that are known to have an association with students' outcomes, in our multilevel models. Appendix 1 lists the variables tested along with descriptions of and sources for these variables.

To determine what factors should be included in the final models, we relied on the statistical significance of the impact each factor has on students' attainment, while holding other factors constant, as well as the stability of the impact across available time series of data[12]. The aim is to identify a stable set of explanatory factors that are shown to have a persistent and significant impact on students' outcomes, to ensure consistent interpretation and comparability of the value added estimates over time and across different cohorts.

Our modelling analysis shows that students' prior test scores, student and school socio-economic status measures, students' Aboriginal and Torres Strait Islander status, whether the student is studying in a class or school that is academically selective (i.e., opportunity classes in primary education or selective schools in secondary education) consistently have a significant predictive relationship with the students' outcomes, regardless which year's data or which cohort of students we are analysing[13]. However, having controlled for these factors in the models, other factors such as school remoteness, school size, proportion of Aboriginal students in a school, school average prior test scores have been shown to have negligible (and sometimes insignificant) impact on students' outcomes, across the multiple years' data files we examined. Under the statistical frameworks used, the effects of gender on student outcomes are found to be negligible when students are in primary and junior secondary years, but significant when they are in the senior secondary years.[14]

Table 1 on the next page lists the variables proposed for inclusion in the DEC VA models.

Two important contextual variables not included in the above table are: (a) a variable indicating whether a student is learning English as an additional language or dialect (EAL/D) and has low English language proficiency (ELP), and (b) a variable indicating whether a student has a confirmed disability. The ELP variable is not included because the department is implementing a new English proficiency assessment framework for EAL/D students in 2014, which will result in new language proficiency classifications of these students. Once new data is available after 2014, further modelling will take place to assess the merit and implications of including an ELP related factor as well as other indicators of EAL/D students' background (such as refugee and new arrival status) in VA models. Similarly, nationally consistent data on students with a disability is expected to be collected for all schools from 2015. Additional modelling will also be undertaken when the new disability data becomes available.

12 Stability of the relative impact each factor has on student learning progress is examined using all available data (i.e. four time periods of data: 2008 to 2010, 2009 to 2011, 2010 to 2012 and 2011 to 2013 - for Year 3-5, 5-7 and 7-9 VA models and 2008 to 2011 and 2009 to 2012 for Year 9-12 VA models).

13 See detailed discussions on the relative impact of the various factors (including gender, single-sex schooling) on student learning progress in Section 5.3.

14 The non-significant gender effect observed when students are in their primary and junior secondary years could be partly due to the fact that the outcome variables used in the NAPLAN-based VA models are composite measures aggregated over reading and numeracy. More detailed discussions concerning the effect of gender on student outcomes in different learning stages is contained in Section 5.3. In order to keep consistent the composition of the senior secondary (9-12) VA models, gender is included as a contextual factor in all three VA 9-12 models.

Table 1: Composition of proposed VA models

<table>
<tr><th rowspan="2">VA measures</th><th rowspan="2">Type of schools applicable</th><th colspan="3">Factors included in VA models</th><th rowspan="2">Dependent variable</th></tr>
<tr><th>Prior ability</th><th>Standard set of contextual factors</th><th>Additional contextual factors</th></tr>
<tr><td>VA Kindergarten to Year 3 (exploratory)</td><td>Primary, Central, Infants</td><td>Student average assessed level across aspects of literacy and numeracy from the Best Start Kindergarten assessment</td><td rowspan="6">Student Aboriginal and Torres Strait Islander status

School SES (Family Occupation and Education Index)

Student SES (based on parental education and occupation)</td><td></td><td>Student average score on reading and numeracy in Year 3 NAPLAN tests</td></tr>
<tr><td>VA Year 3-5</td><td rowspan="3">Primary, Central, Secondary</td><td rowspan="3">Student average score on reading and numeracy in previous NAPLAN tests</td><td></td><td rowspan="3">Student average score on reading and numeracy in later NAPLAN tests</td></tr>
<tr><td>VA Year 5-7</td><td>Student attending an academically selective class (i.e. OC class)</td></tr>
<tr><td>VA Year 7-9</td><td>Student attending a fully academically selective school[15]

Student attending a boys or a girls schools</td></tr>
<tr><td>VA Year 9-12 (TES)</td><td rowspan="3">Secondary, Central</td><td rowspan="3">Student average score on reading and numeracy in previous NAPLAN tests</td><td>Student attending a fully academically selective school

Student attending a boys or a girls schools

Gender</td><td>Student Year 12 Tertiary Entrance Scores (TES)[16] from Higher School Certificate exams (HSC)</td></tr>
<tr><td>supplementary measure: VA Year 9-12 (English)</td><td>Student Year 12 results in English subjects from HSC[17]</td></tr>
<tr><td>supplementary measure: VA Year 9-12 (Maths)</td><td>Student Year 12 results in Maths subjects from HSC</td></tr>
</table>

Note: Descriptions of and sources for the variables are provided in the Appendix 1. When NAPLAN scores are used as the prior and later ability scores, they are derived from students' standardised scores, averaged over reading and numeracy. See Section 4.9 for details of the standardisation process and the rationale behind using scores aggregated over reading and numeracy as the prior and later ability scores.

The next section provides details about how value added measures are estimated.

15 The NSW government system has 17 fully selective high schools with admissions to these schools determined through students' performance on the Selective High School Placement test. There are also four agricultural schools, two of which – James Ruse Agricultural and Hurlstone Agricultural high schools – are fully academically selective. These 19 schools are treated as 'fully selective' schools for the purpose of VA modelling. Two other agricultural high schools – Farrer Memorial and Yanco Agricultural High Schools have a significant boarding section which gives some priority to isolated students. These two agricultural schools are treated as 'not fully selective' schools for the purpose of VA modelling.

16 Tertiary Entrance Score (TES) is an aggregate mark based on a student's best 10 units attempted in HSC exams. Raw examination marks were scaled before they were aggregated. Scaling was considered necessary as a student's position in a course depended on his/her ability and also the abilities of other students in that course. The purpose of scaling was to estimate a student's position in a course if all courses had the same candidature. TES is the basis for determining the ATAR (the Australian Tertiary Admission Rank), which is then used by universities for admission decisions. The department has access to TES scores for research and analysis purposes under the GIPA legislations. More details on the HSC scaling process can be found http://www.uac.edu.au/documents/atar/2012-ScalingReport.pdf

17 Later performance scores used for the VA 9-12 (English) models are calculated from students' scores either on the English Standard 2 Unit or the Advanced English 2 Unit, after raw scores on these two subjects have been scaled to a common scale so that they are comparable. A similar scaling process is also used to equate students' results from different Maths subjects before the scaled scores are used as the later performance scores in the VA 9-12 Maths VA models.

4.4 Calculation of value added measures for schools

To calculate the school VA measures, we first apply the following general equation that illustrates our VA models to estimate model parameters:

$$y_{i,j} = \beta_0 + \beta_1 * X_{1ij} + \ldots + \beta_n * X_{nij} + \beta_{n+1} * W_{1j} + \ldots \beta_{n+k} * W_{kj} + \mu_{0j} + e_{ij} \quad [11]$$

where X_{1ij} to X_{nij} are n student level controlling factors (including student's prior score), W_{1j} to W_{kj} are k school level controlling factors and β_0 is the average performance of all schools, conditional on the student and school factors.

In the context of school effect analysis, μ_{0j} is our primary interest as it represents the contribution each school makes to its students' learning, over and above what can be predicted from student (e.g., background, prior academic achievement) and school characteristics (e.g., demographic and academic composition). μ_{0j} is assumed to have a normal distribution with a mean of zero and variance of $\sigma^2_{\mu 0}$ across the population of schools.

Three steps are involved in estimating μ_{0j}.

The first step is to calculate the raw residual for each student (r_{ij}) in a school (j^{th} school), using model parameter estimates.

$$r_{ij} = y_{i,j} - \widehat{y_{i,j}} \quad [12]$$

where $y_{i,j}$ is the actual performance of student i in j^{th} school

$\widehat{y_{i,j}}$ is the predicted score of this student, given the individual student's background and the school's characteristics

The second step is to calculate the mean of the raw residuals across all students in this school:

$$r_{+j} = \frac{\sum_{i=1}^{n_j} r_{ij}}{n_j} \quad \text{where } n_j \text{ is the number of students in school } j \quad [13]$$

The school effect for the j^{th} school can then be estimated by multiplying this raw average residual r_{+j} by a factor:

$$\widehat{\mu_{0j}} = \frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + \sigma^2_e / n_j} r_{+j} \quad [14]$$

Statistically this factor $\frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + \sigma^2_e / n_j}$ can be regarded as the reliability of the raw residuals and is often referred to as a 'shrinkage factor' because its value always lies between 0 and 1. When the number of students in a school is small, or the within-school unexplained variance (σ^2_e) is relatively large compared to the between-school unexplained variance ($\sigma^2_{\mu 0}$), or both, the shrinkage factor will be considerably less than 1. In either of these two cases, it means there is very little information about the school and the initial school estimate based on the least squares residuals is very imprecise. By applying the shrinkage factor, the estimated school effect is pulled towards the average effect across all schools ($\widehat{\mu_{0j}} \to 0$; $\widehat{\beta_{0j}} \to \beta_0$), as the average effect across all schools is now considered to be a more reliable estimate of the effect for that school (Rasbash et al., 2005).

In essence, applying the shrinkage factor produces an adjusted school effect estimate[18] that is the weighted average of the initial school estimate and the mean estimate across all schools. Use of a shrinkage adjustment is a common technique in the estimation of school or teacher effects, because, without it, the high and the low ends of the distribution of school effects tend to be over represented by small schools just by chance (Johnson et al., 2012). Taking VA 3-5 as an example, our analysis using the latest two time periods of data shows that, without the shrinkage factor, over 80 per cent of schools with VA estimates one standard deviation above or below the mean were schools with matched records less than 10. This is despite the fact that the share of these schools in the population is only 20 per cent. On the other hand, no school with matched records greater than 50 (around 45 per cent of all schools) appeared at the high and low ends of the distribution. After the shrinkage factor was applied, only 11 per cent of schools with VA scores one standard deviation above or below the mean were schools

18 Such adjusted school effect estimates are also referred to in literature as 'Empirical Bayes' estimates.

with matched records less than 10. The corresponding proportion for schools with matched records greater than 50 increased to around 50 per cent.

The use of the shrinkage factor not only improves the precision in the VA measures, but also reduces the volatility in the VA measures from year to year. It also reduces the possibility of schools with a small number of students being falsely identified as performing 'below' average when the school mean performance is unduly influenced by the idiosyncratic performance of few students. This safeguard mechanism is important if the VA measures are used for high-stakes decisions.

4.5 Confidence intervals associated with the VA measures

It is recommended that when reporting a VA measure, the confidence interval around the measure (i.e., the range of the values within which we are statistically confident that the true value of this VA measure lies) be also reported to enable valid comparison of the school effects.

Equation 15 provides the calculation formula for the 95% confidence intervals (CIs) around the estimated school value added measures ($\widehat{\mu_{0j}}$), denoted as [Lower 95% confidence limit, Upper 95% confidence limit]:

$$\left[\widehat{\mu_{0j}} - 1.96 \times \sqrt{\frac{\sigma^2_{\mu 0}\,\sigma^2_e}{\sigma^2_{\mu 0}\,n_j + \sigma^2_e}}\ ,\ \widehat{\mu_{0j}} + 1.96 \times \sqrt{\frac{\sigma^2_{\mu 0}\,\sigma^2_e}{\sigma^2_{\mu 0}\,n_j + \sigma^2_e}}\right] \qquad [15]$$

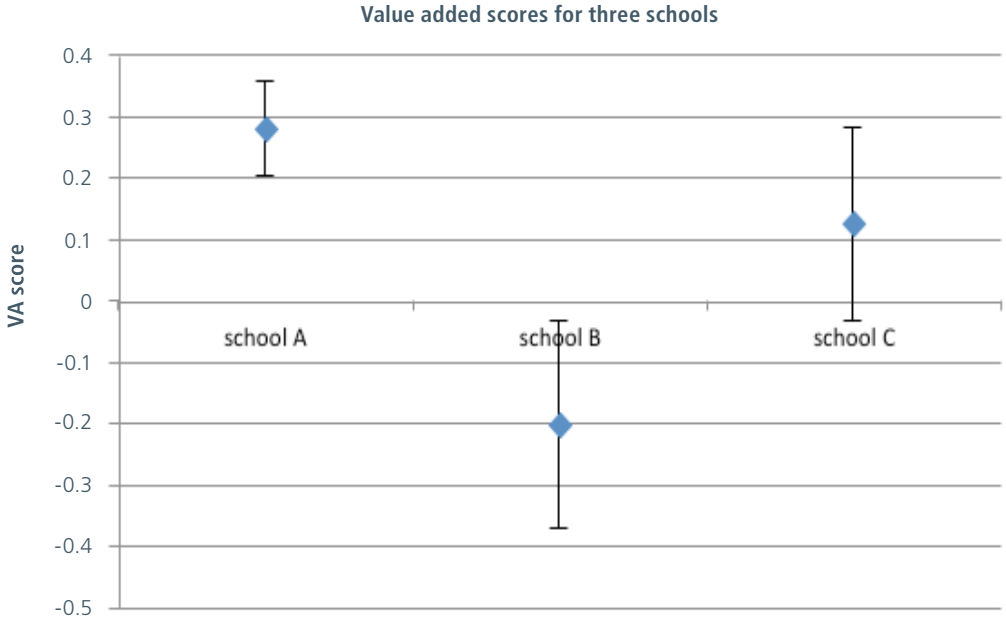
Figure 4 provides an example of how schools' VA scores might be reported in the context of their respective CIs.

As the mean of the VA scores across all schools is zero for each VA measure (e.g., VA 3-5 measure), we can make the following interpretation of a school VA score based on its CI:

- if the lower confidence limit for the school's VA score is greater than 0, the value added by this school can be regarded as statistically above the system average (e.g., school A in Figure 4)
- if the upper confidence limit for the school's VA score is less than 0, the value added by this school can be regarded as statistically below the system average (e.g., school B in Figure 4)
- if the confidence interval straddles the system average of 0, the value added by this school is not statistically different from the system average (e.g., school C in Figure 4).

It is important to stress, in communication to schools, that VA measures are of a relative nature, and that a score of zero does not mean that the school adds no value to students' learning, but only that the school is performing at the system average.

Figure 4:
Reporting VA scores with confidence Intervals



4.6 Use of multiple years of data to generate VA measures

The validity of the VA measures also depends on the extent to which these measures can reliably distinguish the performance of schools (Johnson et al., 2012). In order to demonstrate the extent of the overlapping of the VA scores, Figures 5 and 6 show the VA 7-9 and VA 9-12 scores with confidence intervals, using matched data from the last available time period[19]. The equivalent graphs for VA 3-5 and 5-7 are included in Appendix 2.

Figure 5:

VA 7-9 scores for secondary schools (2011-2013)

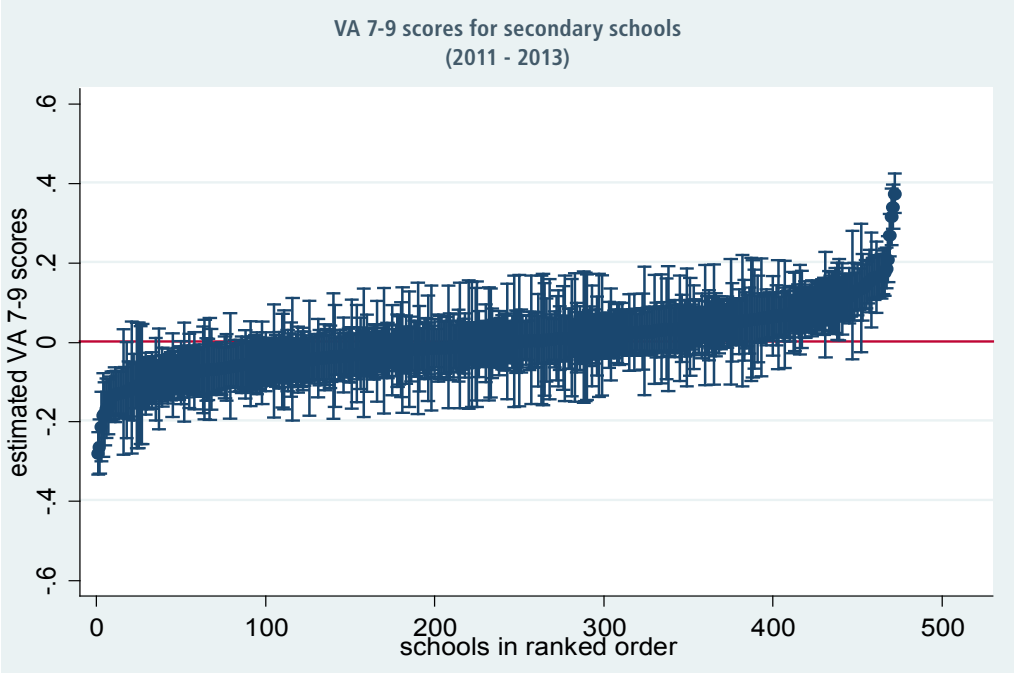
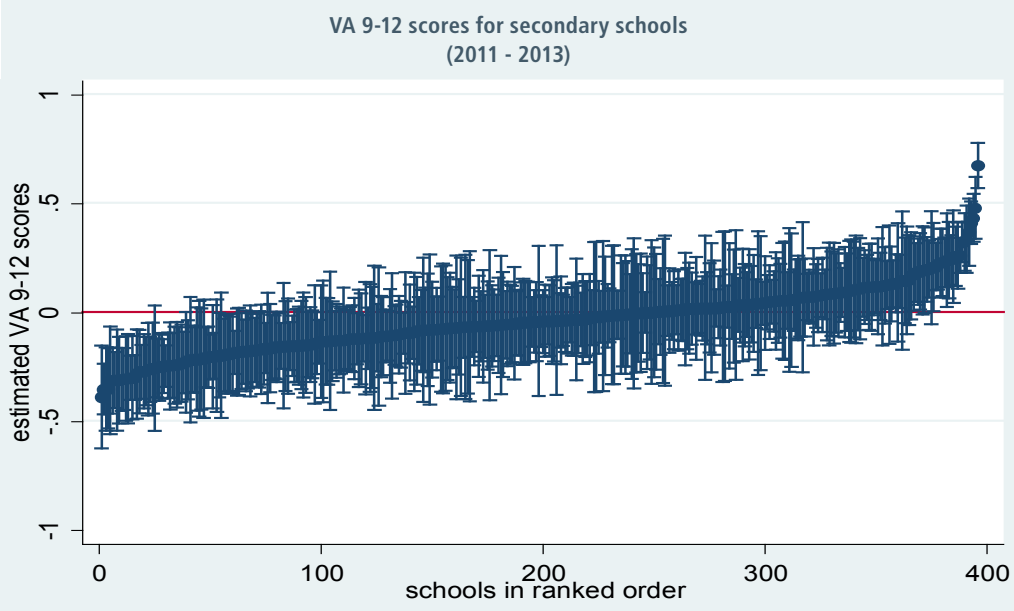


Figure 6:

VA 9 (NAPLAN) to 12 (HSC) for secondary schools (2009-2012)



All graphs show that, for any VA measure, the majority of schools have overlapping confidence intervals and straddle zero, which means their effectiveness cannot be separated from the average with statistical confidence.

A number of factors can impact on the precision of the VA scores. One factor is the reliability of the measures used in estimating school effects. For example the test scores from prior and subsequent tests used in VA models have measurement error themselves, with the reliability of the scores linked to the number of test items.

19 These graphs are based on matched records from the last available single time period; that is matched records from 2011 to 2013 for VA 7-9 and from 2009 to 2012 for VA 9-12.

A second source of error arises from unmeasured transitory factors that impacted on students' test performance for one year. Such factors could include random differences across schools in unmeasured student-level characteristics related to test scores (e.g., some students from one school participated in a one-year literacy program with external researchers) or idiosyncratic unmeasured factors that affected all students' performance in one year (e.g., a fan failing to work in a school hall when a test was administered on a very hot day) (Kane & Staiger, 2002a, 2002b). Since these sources of error are transitory, they not only impact on the precision of a given school's value added score for a particular year, but also are directly reflected in the amount of year-to-year volatility in the VA scores (Schochet & Chiang, 2010).

A common method suggested in the literature to enhance the validity and reliability of the VA scores is to use multiple years' data to estimate VA scores, rather than using just one year or one time period's data (e.g., Johnson et al., 2012; Loeb & Candelaria, 2013; McCaffrey et al., 2009; Schochet & Chiang, 2010). VA measures based on multiple years or time periods are less prone to "statistical noise" (i.e., random errors), and are more likely to reflect the persistent differences in school performance, which is what is intended to be measured. Studies (Godhaber & Hansen, 2010; McCaffrey et al., 2009) investigating teacher value added measures show that using one year of data does only a modest job of predicting teachers' future value added, but using multiple years of data reduces sampling error and improves the accuracy of the prediction.

Our modelling exercise examined the impact of pooling data from two time periods or three time periods on the reliability of the resulting VA 3-5 estimates.

Table 2:

Impact of pooling data across multiple time periods on the reliability of VA 3-5 estimates

VA 3-5	1 Time Period 2011-2013	2 Time Periods 2010-2012 2011-2013	3 Time Periods 2009-2011 2010-2012 2011-2013
% of schools that can be reliably discriminated from the average	17%	22%	26%
Mean standard error	0.09	0.08	0.06

As expected, our analysis shows that the use of multiple time periods' data results in a decrease in the average standard error[20] of VA scores across all schools and an improvement in the number of schools that can be confidently identified as performing above or below the average, for every VA measure we examined. For example, for the VA 3-5 measure (Table 2), the average standard error of the VA estimates decreased from 0.09 when using only one year's data, to 0.06 when using data pooled from the last three time periods. Correspondingly, the proportion of schools identified as performing significantly above or below the average has increased from 17 per cent when one time period of data is used to 22 per cent and 26 per cent when two or three time periods of data are used respectively.

The decision to use two or three time periods of data to estimate VA scores relies on the balance that needs to be struck between two aspects of measure quality – the currency and reliability of the VA information. While VA estimates based on three time periods of data have the lowest mean standard error amongst the three sets of estimates, they encompass growth information spanning over five years which may not be as useful to schools as the other two sets for diagnostic and self-improvement purposes. In addition, true variability in school performance from year to year is also masked to a great extent when VA scores for a given year are based on the last three time periods' of data, as they share two thirds of the underlying data with the VA scores for the preceding year.

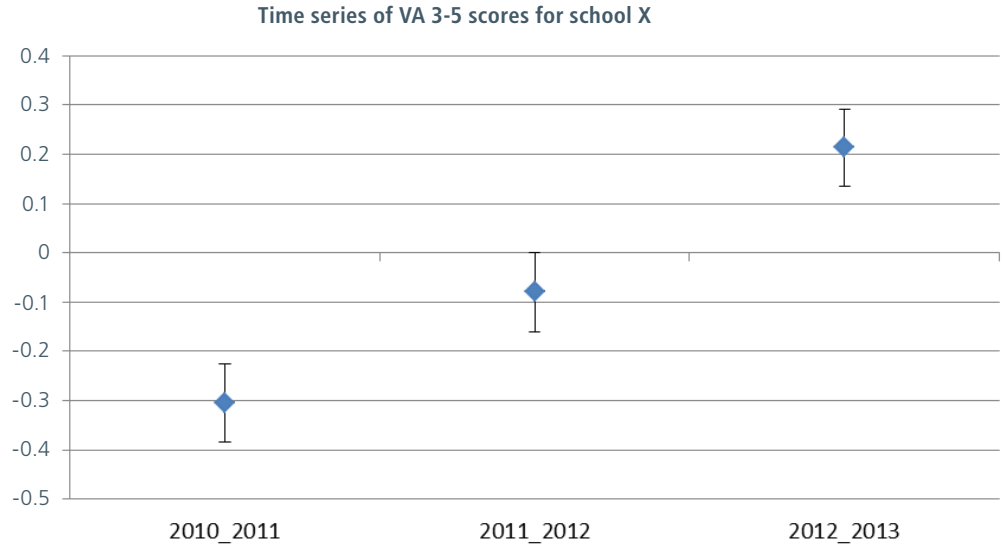
20 The relationship between the confidence interval and the standard error of an estimated VA score is given at Equation 15, where the standard error is estimated as $\sqrt{\frac{\sigma^2_{\mu 0}\,\sigma^2_e}{\sigma^2_{\mu 0}\,n_j + \sigma^2_e}}$

After having considered the various aspects of measure quality, our recommendation is to base VA estimates for a given year on the last two consecutive periods of data, since such estimates can detect performance differences with greater validity and reliability than those based on one time period's data, while still retaining a reasonable level of currency in the VA information for them to be useful to schools for self-improvement purposes. It is also proposed that the reporting of VA measures to schools includes the trend of VA scores over time (an example of such reporting using a real school's data is shown in Figure 7), to help schools identify patterns in school performance.

Figure 7:

Example of reporting time series of VA scores for a school

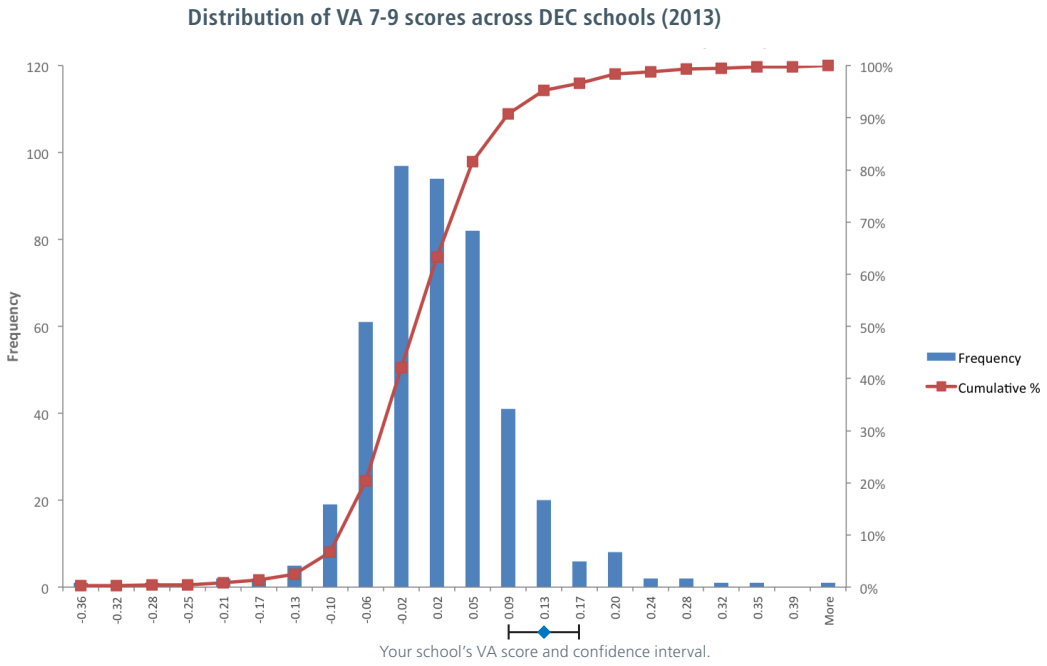
Note: VA measures reported in this figure are based on two consecutive time periods. As an example, VA estimates '2012_2013' refer to those produced based on matched records from 2010 to 2012 and 2011 to 2013.



If a single VA score for a given period of time is to be reported, it is suggested that the score is reported in the context of the distribution of VA scores across all schools (an example is given at Figure 8). This helps accurate interpretation of VA scores, which are inherently of a relative nature.

Figure 8:

Reporting of a VA score for a school, relative to other schools



4.7 Testing of the random slope models

4.7.1 Variability in the slope for student SES across schools

The proposed VA models for NSW government schools constrain the relationships between contextual factors (such as students' SES) and students' achievements to be invariant across schools. In other words, the respective coefficients associated with the factors are assumed to be the same for all schools, reflecting the underlying conceptual idea that the VA models ought to adjust for the average impact these factors have on students' outcomes across all schools.

However, the proposed models can be extended to investigate whether these relationships do vary across schools in the population. If the variability of the slopes (coefficients) is statistically significant, the finding on its own would be of substantive interest as it can shed light on whether schools are differentially effective with different types of students.

This section explores the question of whether the relationship between student outcome and student SES is constant across schools, having controlled for all other factors. We also explore the impact this modelling choice has on the school effects estimates, by comparing the VA scores from the random slope model to those from the simpler model we proposed, which is the random intercept model.

To demonstrate our findings, the 2011 Year 7 students across all NSW government schools were selected as a sample cohort. Two VA models were developed to examine the learning progress these students made when they reached Year 9 in 2013. The first model – our proposed model, hereafter referred to as Model 1 – is a random intercept model which assumes fixed coefficients for all explanatory variables. The second model (hereafter referred to as Model 2) differs from the first by only one aspect, that is, the slope (denoted as β_{1j} in the following equation) for the student SES (denoted as X_{1ij}) is hypothesised to vary across schools:

$$y_{i,j} = \beta_0 + \beta_{1j} * X_{1ij} + \ldots + \beta_n * X_{nij} + \beta_{n+1} * W_{1j} + \ldots + \beta_{n+k} * W_{kj} + \mu_{0j} + e_{ij} \quad [16]$$

where $\beta_{1j} = \beta_1 + \mu_{1j}$ [17]

μ_{1j} represents the deviation of the slope for the j^{th} school from the average slope of β_1, and is assumed to be randomly distributed with a mean of zero and variance of $\sigma^2_{\mu 1}$, and

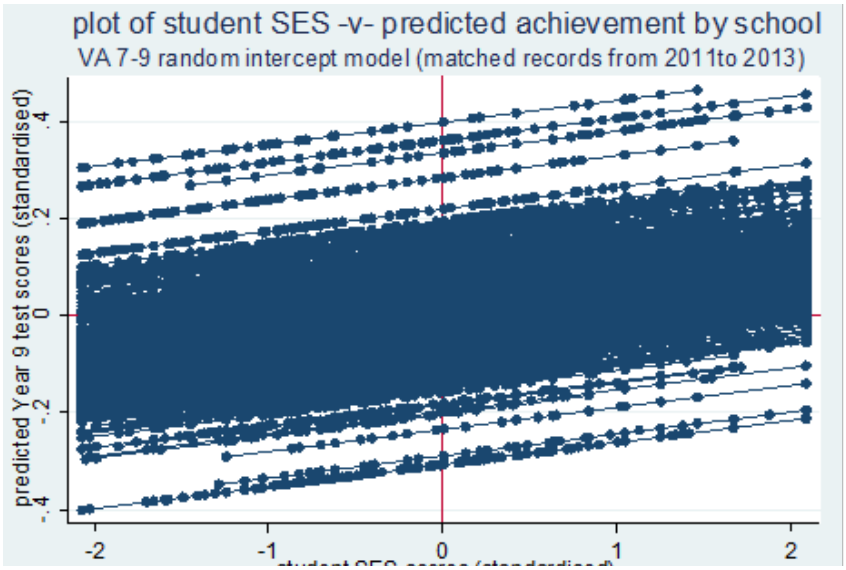
X_{pij} and W_{qj} denote the series of student level (including students' prior test scores) and school level covariates included in our VA models respectively.

Figure 9 shows the predictive relationship between student SES and student outcome for each school, estimated from Model 1, while holding all other variables constant (in this case, by setting their values to their respective grand means)[21]. Each school has an estimated regression line, with each dot on a line representing a student in the school. The vertical axis is the predicted score for each student, and the horizontal axis is the student's SES score – the higher the SES score, the higher the student's socio-economic status. The regression line for a given school is different from the average line in its intercept, by an amount $\widehat{\mu_{0j}}$, which is the estimated school VA score. These lines are parallel to each other because this model assumes the impact of SES on students' outcomes (i.e., the slope) is invariant across schools.

21 Both of the VA models used in this section include the following controlling factors: student SES scores, school FOEI, students' prior test scores averaged across reading and numeracy, and students' ATSI status and school selective status. The first three variables and the dependent variable - Year 9 test scores (averaged across reading and numeracy) are all standardised to a mean of zero and standard deviation of 1. For ease of interpretation, the two dichotomous categorical variables (ATSI status and school selective status) have also been centred around the mean (i.e. the frequency of the respective base category) in the population before multilevel modelling. The regression lines from the two VA models demonstrated in Figures 9 and 10 are those when all explanatory variables in the model (other than the student SES) are held constant – i.e. all set to their respective grand means.

Figure 9:

Plot of student SES and student predicted achievement by school (random intercept only model)

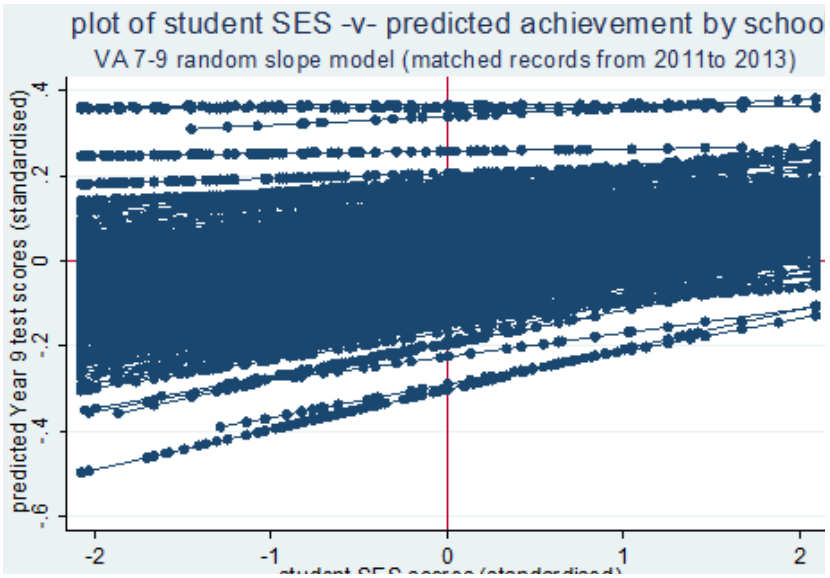


Overall, consistent with other studies, student SES has a significant predictive relationship with the student outcome ($\hat{\beta} = 0.045, \text{se} = 0.003$). Everything else being equal, the higher the student's SES, the higher the predicted test score on the Year 9 NAPLAN tests. For an increase of one standard deviation in the SES score, the predicted test score in Year 9 is expected to increase by a 0.045 standard deviations, after everything else is held constant[22].

When Model 2 is fitted, the likelihood ratio test indicates that the variability in the SES-achievement slope across schools is statistically significant[23] and that the more complex slope varying model (Model 2) is a better fit to the data than the simpler model (Model 1). The average slope is estimated to be 0.046, which is close to the coefficient estimated from Model 1. Assuming a normal distribution, 95 per cent of the schools are expected to have a slope between 0 [a flat line] and 0.10. This variability can be visually illustrated in Figure 10, where all other variables have been set to their respective grand means.

Figure 10:

Plot of student SES and student predicted achievement by school (random slope model)



Two observations can be made from Figure 10. First, it appears that schools with lower intercepts have steeper regression lines. This is further confirmed by statistical analysis showing that the intercepts are negatively correlated with the relative slopes (μ_{1j}) at -0.37 (this relationship is plotted in Figure 11). This finding is of substantive interest,

22 See Section 5.3 for more discussion about the relative impact of contextual factors on student learning progress.

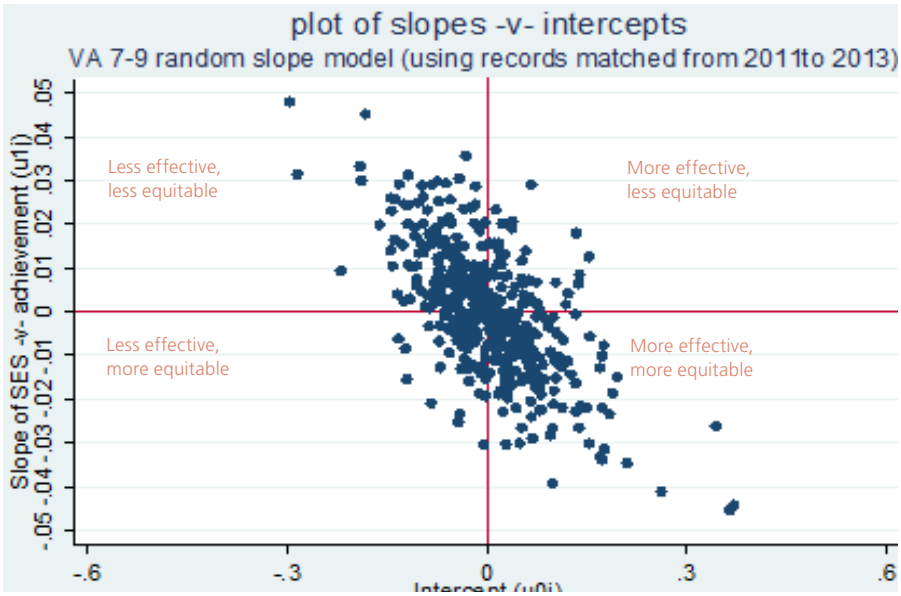
23 The log-likelihood value decreased from 48809.38 for Model 1 to 48774.55 for Model 2. The change in the -2log-likelihood value is 34.8, which has a chi-squared distribution with 2 degrees of freedom. Therefore the null hypothesis that the extra two parameters (i.e. the variation in the slopes and the covariance between the school intercepts and slopes) are simultaneously equal to 0, is rejected and we can conclude that Model 2 is a more elaborate model that better fits the data than Model 1.

as it indicates that more effective schools also tend to be more "equitable", with a weaker SES-achievement relationship. Schools that are relatively more effective and more equitable – that is with higher adjusted average achievement (positive values of μ_{0j}) and weaker SES effects (negative values of μ_{1j}) are found in the lower right hand quadrant of Figure 11. Schools with the opposite characteristics (less effective and less equitable) can be found in the upper left quadrant. Ideally we would like all schools to be effective as well as equitable, so the schools in the lower right hand quadrant could be targeted for follow up investigation to find out what these schools do differently from the rest.

Figure 11:

Scatterplot of intercepts -v- slopes (from VA7-9 random slope model)

Note: The vertical axis shows the slope of student SES versus achievement for each school, relative to the average slope in the population. Similarly, the horizontal axis shows the intercept for each school, relative to the average intercept in the population, which is close to zero in this case.

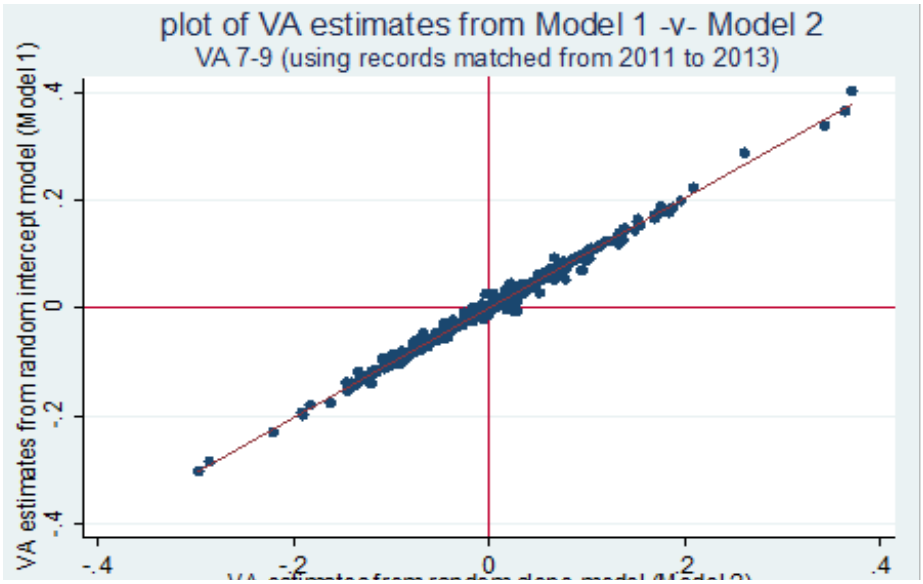


The second observation from Figure 10 is that the difference in the predicted test scores of students from the lowest SES background attending the most effective versus the least effective school, is around 0.8 standard deviations (when everything else is held constant). This difference (or educational benefit) is considerably greater than that for high SES students. Students of the highest SES background achieve 0.5 standard deviations more attending the most effective than the least effective schools, over a two year period. This suggests school effectiveness makes the greatest impact on low SES students, that is, low SES students benefit more from improved school effectiveness than high SES students.

Though the random slope model is a more elaborate model that fits the data better than the simpler fixed slope model, its impact on the estimated VA scores is negligible. This is demonstrated in Figure 12 where the VA scores

Figure 12:

Scatterplot of VA 7-9 estimates from Model 1 (proposed model) and Model 2 (random slope model)



from the two models are shown to align extremely well with the correlation between the two sets of scores close to 1. This is largely attributable to the fact that, although the variability in the slope is statistically significant, in practical terms it is still very small, with slopes for 95% of the schools falling in the range between 0 and 0.10. Additionally the variability in the slopes is much less than the variability in the intercepts, with the variance in the intercepts between schools over 10 times greater than the variance in the slopes between schools.

4.7.2 Variability in the slope for prior achievement across schools

Variability in the slope for prior scores is also investigated by varying our proposed models to include a random component which represents the deviation of each school's slope for prior ability level from the average slope in the population[24].Analysis shows similar results to those reported in the preceding section. For instance, allowing slopes for prior scores to vary across schools did not have a significant impact on the VA 7-9 scores[25]. The more complex approach produced similar VA scores to those from the simpler approach for most schools, with an overall correlation of 0.98 between the two sets of VA scores.

4.7.3 Conclusion

The simpler modelling technique – constraining the coefficients to be invariant across all schools – is favoured over the more complex technique for a number of reasons. First, even though the more complex technique results in more elaborate models that fit the data better, the two modelling techniques produced similar VA scores. The more complex approach is also more difficult to explain to stakeholders. Lastly from a conceptual point of view, it seems educationally defensible to adjust for the average impact of important contextual factors on student outcomes, rather than allowing for this impact to vary across schools, which may be conceived as 'relieving' schools' responsibilities to close gaps between equity groups.

While the random slope modelling is not proposed for VA estimations, it is a valuable technique that can provide useful diagnostic information to help schools and the system to identify areas for further improvement. It is important to stress that this technique, like any other VA techniques, does not tell us why some schools are performing better or worse than expected. However, it does help us to identify schools that have particular performance patterns, so that more in-depth follow up investigations can be carried out. It is through these investigations that we can identify ground level policies and practices that make a difference in schools' ability to improve students' learning outcomes.

4.8 Modelling of the non-linear effects of prior achievement

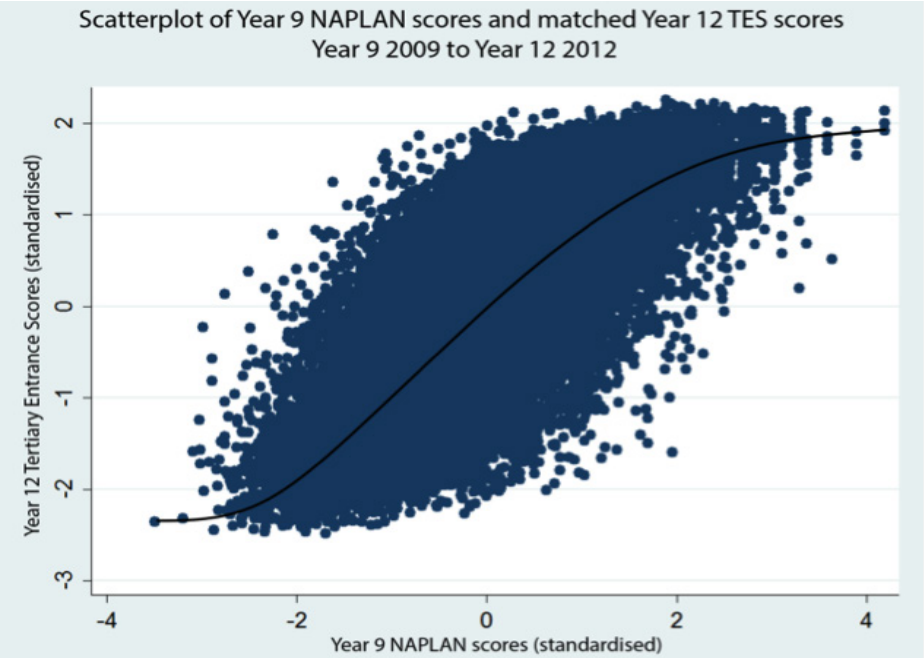
Previous DEC analysis on student growth using NAPLAN test scores has consistently demonstrated that students with high and low prior achievement levels have different trajectories than students who score near the middle of the distribution. For an example, based on Year 9 students' test scores in 2009 matched to their Year 12 results in 2012, Figure 13 shows that the relationship between prior and later achievement scores is not linear. The horizontal axis of the figure represents students' (standardised) average NAPLAN scores over reading and numeracy in Year 9 and the vertical axis represents their (standardised) Tertiary Entrance Scores estimated from their performance on the Higher School Certificate exams in Year 12. It is clear that the relationship tapers off at the high end of the prior achievement scale. Students with high prior achievement scores do not gain as much as students who have average prior scores. Conversely, students with low prior scores make greater gain relative to those who have average prior scores.

Such a non-linear relationship also exists in other VA datasets (e.g., matched records from Year 3 to Year 5, from Year 5 to Year 7 and from Year 7 to Year 9)[26] and in other educational system data (Johnson et al., 2012); and is thought

24 This analysis is based on testing the variability of the linear relationship between student prior and later test scores across schools. The nature of this relationship, whether linear or non-linear, is further discussed in the next section.
25 This analysis used the matched data from 2011 Year 7 to 2013 Year 9.
26 See the coefficients (with statistical significance) reported for the non-linear effects of prior achievement included in the VA models in Section 5.3.

Figure 13:

Relationship between Year 9 NAPLAN scores and matched Year 12 TES scores



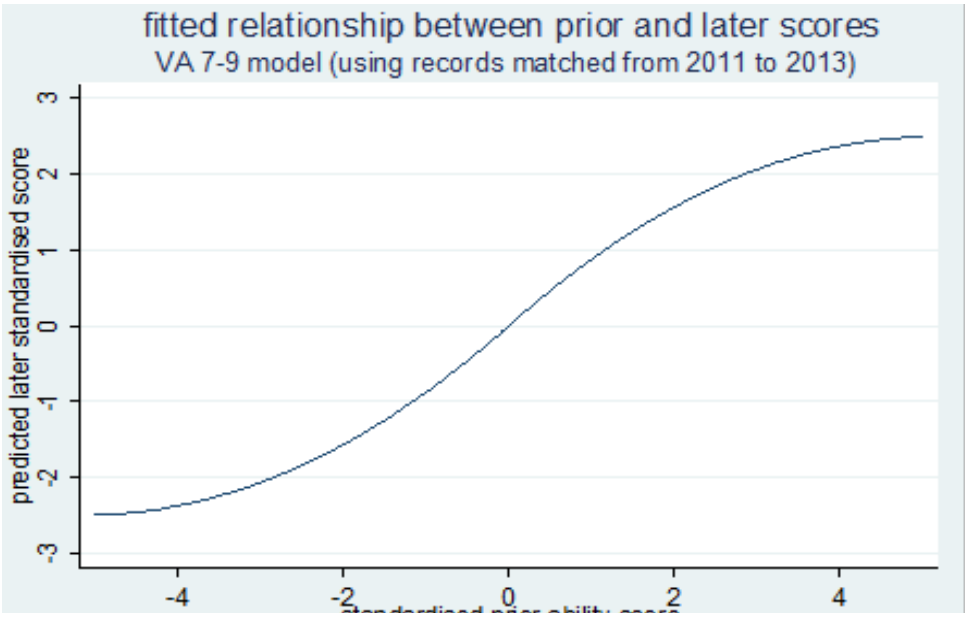
to be related to measurement issues associated with standardised tests (e.g., "floor" or "ceiling" effects) (Johnson et al., 2012). High performing students may appear to have made less gain than average performing students because the tests are not sensitive to their real proficiency levels and the progress they are making. Similarly, low performing students may appear to have made greater gain because the tests are not designed to measure the proficiency levels that are substantially below national minimum standards with a high level of precision.

Such differences in the expected learning trajectories, which may be caused by heterogeneity in the degree of measurement error across the distribution, need to be adequately addressed in the VA models so that the estimated school effects do not disadvantage or advantage schools who enrol a large number of high or low performing students.

Our modelling shows that the differential trajectories can be modelled using a quadratic polynomial function illustrated in Figure 14. With the horizontal axis representing the distribution of student (standardised) prior test scores, and the vertical axis the corresponding predicted scores from the subsequent test for the same students, the curve accords with the above-mentioned pattern observed in the data. Statistical analysis shows that the addition

Figure 14:

Fitted non-linear relationship between prior and later test scores



of such a quadratic function to a base model, which includes only a linear function of the starting ability, makes a significant improvement to the fit of the model to the data[27]. Further analysis also shows that, while such an addition does not have any significant impact on the VA estimates produced[28], it does result in a very small number of schools changing their VA estimates from 'at average' to 'significantly above or below average' or vice versa. Since DEC schools are familiar with the non-linear predictive relationship between prior and later test scores – as it is visible in the SMART package – it is recommended that a quadratic polynomial function of prior scores be included in the proposed VA models to further improve the validity of VA estimates.

4.9 Other features of the proposed VA models and measures

Profile of VA measures for each school

As discussed earlier, our proposed VA models are estimated separately for each cohort (e.g., Year 3 students matched to their results when in Year 5) and for each given time period (i.e., for every two consecutive time periods). This results in a profile of different VA measures, each with its own time series, estimated for each school (see Table 3). While VA measures for different cohorts of students might be combined into a composite measure for each school, this approach is not favoured because DEC analysis and other studies (Nuttall et al., 1989; Sammons, Nuttall & Cuttance, 1993) suggest schools could be differentially effective with different cohorts of students. So that the VA measures can be most useful to schools for diagnostic and local improvement purposes, the profile-based approach in presenting the VA information is recommended.

Table 3:

Proposed profile of VA measures for each school, developed from the modelling exercise

Note: this table lists all the VA measures that have been generated through our modelling exercise, which can be updated on an annual basis when new test data is received.

VA measures	Type of schools applicable	Time series available	Underlying data source (pooled across two consecutive periods where possible)
VA Kindergarten to Year 3 (exploratory)	Primary, Central, Infants	2013	2010 Kindergarten matched to 2013 Year 3
VA Year 3-5	Primary, Central, Secondary	2011	2009 to 2011, 2008 to 2010
VA Year 5-7		2012	2010 to 2012, 2009 to 2011
VA Year 7-9		2013	2011 to 2013, 2010 to 2012
VA Year 9-12 (Tertiary Entrance Scores) &	Secondary, Central	2012	2009 to 2012, 2008 to 2011
VA Year 9-12 (English)			
VA Year 9-12 (Maths)			

Variable standardisation

Each VA model is estimated without the constant term and after all continuous variables including the later test scores have been standardised[29]. This process improves the precision (Johnson et al., 2012) as well as the interpretability of the results, as the VA estimates therefore reflect the school effects relative to the average for all schools, and can be interpreted in units of standard deviations in the later test scores.

27 For example, using the matched data from Year 7 (2011) to Year 9 (2013), comparing a VA model fitted with a quadratic polynomial function of prior scores to a base model with the same controlling factors but only a linear function of prior scores, the change in the -2log-likelihood value is 449.42, which has a chi-squared distribution with 1 degree of freedom. The change is highly significant, confirming that the more sophisticated model is a model that better fits the data.

28 VA estimates produced from the more sophisticated model, which includes an adjustment for the non-linear relationship, and the simpler model, which includes only a linear relationship, align very closely, with an overall correlation at 0.98.

29 Standardising a variable involves, for each value, subtracting the average value from it and then dividing it by the standard deviation of the values. For the prior and later NAPLAN performance scores used in the models, they are derived from first standardising each student's test score within the relevant government student cohort for that test, and then averaging the standardised scores over reading and numeracy. The aggregate scores are then re-standardised to ensure the mean and standard deviation of the outcome variable used in each model are 0 and 1 respectively.

Students included in the VA analysis

For all VA measures, only matched students who were attending the same school at the two consecutive testing times contribute to the estimation of the relevant VA measures for that school, except in the following cases:

- for the K-3 measures, the value added data for those students who came from infants schools (K-2 schools) are credited to the relevant infants school, since 90 per cent of the period between Kindergarten and the Year 3 NAPLAN testing time occurred in the infants school;
- where students moved from one primary school to an opportunity class in another school at the beginning of Year 5, value added Year 3 to Year 5 data for that student is credited to the first primary school where the student was located for most of the period between the Year 3 and Year 5 testing times;
- for the Year 5 to 7 VA measures, the value added data for the matched records are credited to the primary schools since 90 per cent of the period between the two test times occurred in the primary schools.

Further discussion about how students who moved schools during a measurement period could be included in the future versions of the VA models is provided in Section 6. The treatment of this group of students is particularly important for the validity of the VA 9-12 estimates as students may change schools when they transition from junior to senior secondary education, and this further exacerbates the problem associated with the attrition of students from Year 9 to Year 12 for the VA 9-12 measures.

Aggregate test scores used in VA models

All the proposed VA measures use aggregate test scores, therefore they are not subject specific[30]. In VA 3-5, 5-7 and 7-9 models, both the later and the prior test scores are the (standardised) average reported scores over the reading and numeracy NAPLAN tests for each student. For the main VA 9-12 measure, the prior score is the (standardised) average reported score over the Year 9 NAPLAN reading and numeracy tests for each student, and the later score is the same student's (standardised) Year 12 Tertiary Entrance Score, a weighted score across the best 10 HSC units the student attempted.

The choice of using aggregate test scores as proxies of prior and later achievement levels is partly due to evidence that school effectiveness tends to 'spill over' across tests (Deming, 2014). For example, analysis of separate VA measures for 9-12 English and 9-12 Mathematics shows that the correlation between the two value added measures for the same time period is about 0.6. Another reason for using aggregate scores is that all assessments measure students' subject specific skills and understandings with some degree of error and test scores aggregated over multiple tests are more reliable indicators of students' overall achievement level at a particular time than a single test score. This not only improves the precision of VA estimates but also helps mitigate potential issues related to test floor and ceiling effects (Johnson et al., 2012).

The next section discusses the results from an analysis of the VA estimates developed for NSW government schools.

30 The only exception are the two supplementary subject specific VA 9-12 measures – VA9 (NAPLAN)-12 (HSC English) and VA9 (NAPLAN)-12 (HSC) Maths – developed for estimating value added by schools to students' learning on the English and Maths subjects from Year 9 to Year 12, respectively.

5 Discussion of modelling results

5.1 Distribution of school effects estimates

Table 4 provides the results of the school effects estimated for VA 3-5, 5-7 and 7-9 and 9-12, using the last two pooled time periods of data.

Table 4: Results from VA analysis

VA measures	Students	Schools	Percentage of schools that can be discriminated from the average with confidence (95%CI)	Mean standard error	Highest VA score	Lowest VA score	90th percentile	10th percentile
VA 3-5	90,846	1,641	22%	0.075	0.59	-0.48	0.12	-0.13
VA 5-7	83,285	1,636	21%	0.067	0.42	-0.54	0.09	-0.12
VA 7-9	77,723	446	35%	0.032	0.43	-0.36	0.09	-0.08
VA 9-12 (TES)	47,851	440	20%	0.079	0.40	-0.32	0.07	-0.16

Table 4 shows that, using a 95 percent confidence interval, around 20 per cent (for VA 9-12) to 35 per cent of schools (for VA 7-9) can be distinguished from the system average. In addition, the spread of the value added estimates (in standard deviation units) varies across the cohorts of students under examination. At the extremes, a school with the highest VA 3-5 score is estimated to raise student achievement on the later tests (i.e., students' average performance on Year 5 NAPLAN reading and numeracy tests) by 0.6 standard deviations compared to the average school in the system. The equivalent additional value a top performing school brings to its junior secondary students' learning, relative to the average school in the system, is 0.43 standard deviations on the later achievement (i.e., students' average performance on Year 9 NAPLAN reading and numeracy tests).

It is worthwhile noting that, across the VA measures very few schools (less than 2 per cent of primary schools and less than 1 per cent of secondary schools) have value added scores that are greater than 0.3 or lower than -0.3. For this reason, VA scores at the 90th percentile and 10th percentile are also reported in Table 4. As compared to a school at the 10th percentile, a school at the 90th percentile improves a Year 3 student's achievement on Year 5 NAPLAN results by an additional 0.25 standard deviations, and a Year 7 student on Year 9 NAPLAN tests by 0.17 standard deviations. When these differences are expressed in terms of learning gain, it is estimated that an average Year 3 student attending a 90th percentile school is approximately five months ahead of a similar student attending a 10th percentile school by the time they reach Year 5. Similarly, an average Year 7 student attending a 90th percentile school is around nine months ahead of a similar student attending a 10th percentile school by the time they reach Year 9. The relative learning gain is larger for secondary students as the typical gain from Year 7 to Year 9 is about half the typical gain from Year 3 to Year 5, so the same amount of improvement in NAPLAN reported scores would equate to greater relative gain for secondary students than for primary students.

When interpreting the results contained in Table 4, it is noted that the level of discrimination and the spread of VA scores associated with the VA 9-12 measure is impacted by the differential rates across schools of students leaving before completing Year 12. These students do not contribute to the estimation of the VA 9-12(TES) measure and they are more likely to be from a low SES background. It is recommended that DEC explores non-test outcome based value added measures such as those gauging the contributions schools make to retain students to Year 12 to add to the profile of VA measures, so that schools' success can be estimated on a more balanced and complete basis.

5.2 Stability in VA estimates

Stability in VA estimates from year to year is also examined. Bearing in mind that VA estimates for each year are calculated by pooling data from the last two time periods, Table 5 shows that a VA measure for one year correlates with the same measure for the next year in a range from 0.7 to 0.8. Additional analyses show that, if schools are classified into three levels (at average, significantly above average, significantly below average) based on the 95% confidence intervals around their VA estimates, 75 to 80 per cent of schools would retain the same classification from one year to the next, and the rest (20 to 25 per cent) of schools would change by one classification level (e.g., from average to significantly above or below average). Only a tiny number of schools would change by two classification levels (i.e., from significantly below to significantly above or vice versa) from one year to the next. This level of inter-year variability is considered satisfactory for the VA measures. For system improvement purposes, the number of schools identified as consistently performing above average across years or having experienced positive changes on VA in a given year (i.e., those moved up a classification level from the previous year) is regarded large enough for DEC to conduct follow-up investigations as to what drives these schools' success.

A moderate to strong relationship between VA estimates across time is to be expected since the adjacent year's VA estimates share one year's matched records. If only one year's data were used to produce VA estimates, the year-to-year correlation would range from 0.3 to 0.6, and the classification for many more schools would change from year to year. This level of stability is consistent with existing studies on teacher and school effects (focusing on grade-level teams), which also reported weak to moderate year-to-year correlation (0.2 to 0.6) based on one year's data (Goldhaber & Hansen, 2008; Kane & Staiger, 2002b; McCaffrey et al., 2009).

Our stability analysis also shows that VA estimates are more stable for secondary cohorts than for primary cohorts, irrespective of how many years of data are used to calculate the VA estimates. For example, Table 5 shows that the correlation between two VA estimates that do not share matched records is higher for VA 7-9 than for VA 3-5. Whether this pattern is caused by a greater level of true variation in value added by schools to their middle primary students over time than to other cohorts of students, or it is simply due to a greater level of noise in the VA 3-5 data, needs to be investigated by future studies.

Table 5: Correlations between VA estimates over time

	2013 VA vs 2012 VA	2012 VA vs 2011 VA	2013 VA vs 2011 VA
Underlying matched records used	Pooled records from 2010-12 & 2011-2013 vs Pooled records from 2009-11 & 2010-2012	Pooled records from 2009-11 & 2010-2012 vs Pooled records from 2008-10 & 2009-2011	Pooled records from 2010-12 & 2011-2013 vs Pooled records from 2008-10 & 2009-2011
	Share one year matched records	Share one year matched records	No shared records
VA 3-5	0.68	0.64	0.27
VA 5-7	0.76	0.77	0.5
VA 7-9	0.8	0.76	0.56

5.3 Relative impact of contextual factors on student learning progress

VA analysis can also shed light on the relative importance of contextual factors in explaining the variation in students' learning progress.

Tables 6 and 7 report the coefficients from primary and secondary VA models for the various factors, using data pooled across the available latest two time periods. Note that, for all VA models, dependent variables and all continuous explanatory variables were standardised prior to multilevel modelling.

Table 6:

Modelling results from primary VA 3-5 and 5-7 models (using data from the latest two time periods)

Note 1: For ease of interpretation, school SES scores are FOEI scores multiplied by (-1), so that higher SES scores reflect higher SES.

Note 2: Opportunity Classes are only available in Year 5 and Year 6, therefore any unique educational advantage associated with attending an OC is taken into account in all VA 5-7 models but not in VA 3-5 models, since only 10 per cent of the period between Year 3 and Year 5 NAPLAN test times occurs in opportunity classes, if a student is selected into such a class at the beginning of Year 5.

Later year test scores (standardised)	VA 3-5		VA 5-7	
Fixed effects	Coefficient	Standard error	Coefficient	Standard error
Student characteristics				
Prior achievement score (standardised)	0.86	0.00	0.95	0.00
Adjustment for non-linear relationship between prior and later achievement score (standardised)	-0.09	0.00	-0.16	0.00
Aboriginal & Torres Strait Islander (A&TSI) (vs non A&TSI as reference category)	-0.07	0.01	-0.07	0.01
Student SES score (standardised)	0.08	0.00	0.07	0.00
Attending an Opportunity Class (OC) (vs not attending an OC as a reference category)	--	--	0.36	0.01
School characteristics				
School SES score (standardised)	0.06	0.00	0.04	0.00
R-squared using Snijders/Bosker (1994, 1999)				
Proportion of variance explained at Level 1 (student)	72%		78%	
Proportion of variance explained at Level 2 (school)	85%		88%	

Table 7:

Modelling results from secondary VA 7-9 and 9-12(TES) models (using data from the latest two time periods)

Note: For ease of interpretation, school SES scores are FOEI scores multiplied by (-1), so that higher SES scores reflect higher SES

Later year test scores (standardised)	VA 7-9		VA 9-12(TES)	
Fixed effects	coefficient	standard error	coefficient	standard error
Student characteristics				
prior achievement score (standardised)	0.98	0.00	0.77	0.01
adjustment for non-linear relationship between prior and later achievement score (standardised)	-0.16	0.00	-0.17	0.01
Aboriginal & Torres Strait Islander (A&TSI) (vs non A&TSI as reference category)	-0.05	0.01	-0.17	0.03
student SES score (standardised)	0.04	0.00	0.08	0.00
boys (vs girls as a reference category)	--	--	-0.25	0.01
School characteristics				
fully selective school (vs comprehensive school as a reference category)	0.25	0.02	0.04	0.02
boys-only school (vs co-educational school)	0.08	0.02	0.23	0.04
girls-only school (vs co-educational school)	0.06	0.02	0.22	0.04
school SES score (standardised)	0.03	0.01	0.13	0.01
R-squared using Snijders/Bosker (1994, 1999)				
Proportion of variance explained at Level 1 (student)	84%		55%	
Proportion of variance explained at Level 2 (school)	96%		83%	

Prior achievement scores

As expected, amongst all factors, prior achievement scores have the greatest impact on students' later scores. Everything else being equal, one standard deviation increase in prior scores equates to at least 0.8 standard deviations increase in the later scores, across all VA measures[31].

Compared to a null model which does not include any explanatory variables, adding prior test scores (and a polynomial function of prior scores) alone would reduce the unexplained variance at the student level by around 40 per cent for the VA 9-12(TES) models and around 67 per cent to 75 per cent for other NAPLAN based VA models. In comparison, inclusion of other student level variables would reduce the unexplained variance at the student level by about 5 to 15 per cent, depending on the VA models tested.

Student and school background factors

Since the prior test scores have captured the impact of student level contextual factors such as ATSI and SES to a large extent, the residual impact of the individual contextual factors on students' learning is relatively small (mostly less than 0.1 standard deviation). However, relative disadvantage could accumulate and become significant when a student experiences multiple aspects of disadvantage. For example, compare the learning progress of two students

31 This is based on results from VA models which have the same model specifications as those in Tables 6 & 7 but exclude an adjustment for the non-linear relationship between prior and later scores, for ease of interpretation of the impact of the prior scores.

who, apart from family and school SES and student's Aboriginality, are identical on all other aspects. One is a typical non-Aboriginal Year 3 student, who had an average parental SES background and attended a school of average SES. Another is an Aboriginal student who achieved the same prior score in Year 3 as the non-Aboriginal student, but was of low SES background (being one standard deviation below average) and attended a low SES school (whose SES score was one standard deviation below average school SES). The latter is estimated to underperform the former by 0.2 standard deviations by the time they reach Year 5. The disadvantage is the most pronounced for senior secondary Aboriginal low SES students. It is estimated that, by the time these students have reached Year 12, they underperform typical non-Aboriginal students by 0.38 standard deviations on HSC exams, even if they had the same prior achievement level in Year 9 as a typical non-Aboriginal student.

Clustering of high performing students

The effect of selecting high achieving students in a class or in a school is also a persistent factor influencing student outcomes, even after variation in student prior achievement scores and other background factors have been taken into account. Studying with other students at similarly high achievement levels improves student performance by 0.36 standard deviations for upper primary students and by 0.25 standard deviations for junior secondary students. While this effect is considerably smaller for senior secondary students, it could be due to the fact that about 40 per cent of students, mostly low performing students, either drop out of school before completing Year 12, move to a non-government school, or undertake an HSC course of study that is not eligible for the TES calculation. They are therefore not included in the VA 9-12(TES) models, and this is likely to have contributed to a reduced effect for 'clustering'.

It is worthwhile noting that the effect discussed above is the effect on individuals placed in a cluster of high performing students. This effect is different from the net effect the streaming practice (i.e., streaming students based on academic achievements) might have on the system as a whole, which would need to include the effect of such a practice on the other students who are not selected into selective classes or schools.

Gender effect

Our modelling exercise also examined the effect of gender on student outcomes in different learning stages. We find that, having adjusted for prior scores and important contextual factors, the gender effect is negligible in all VA models except in the VA 9-12(TES) models, where the effect size is 0.25 standard deviation, favouring girls. Further analysis using separate VA 9-12 (English) and VA 9-12 (Maths) measures show that, everything else being equal, girls outperform boys of similar background and with similar prior achievement level, by 0.4 standard deviations in Year 12 English results, but this advantage is non-existent when it comes to their performance in Maths subjects[32]. Thus the gender effect reported in Table 7 for Year 9-12 (i.e., 0.25 standard deviations) reflects the relative effect of gender on students' overall performance on HSC subjects, given Tertiary Entrance Scores – used as the outcome variable in VA 9-12 models – are weighted average scores over multiple HSC subjects (including English) a student attempted.

Single-sex schooling

A final note about the results reported in Table 7 relates to the estimated effect associated with single-sex schooling. The merit of educating students in sex segregated schools has been debated and researched both in Australia and abroad for a long time (e.g., Baker, Riordan & Schaub, 1995; OECD, 2006; Salomone, 2003; Streitmatter, 2002). However, the overall evidence in this debate remains inconclusive (Sikora, 2013). While it is not within the scope of this paper to discuss and debate the advantages and disadvantages of single-sex schooling, we nonetheless note that our VA analysis confirmed positive effects of single sex-schooling in NSW government system[33]. After the variation between schools in student intake policies and other student characteristics were taken into account, the effect associated with single-sex schooling ranged from 0.08 standard deviations for junior secondary students

32 One explanation that a significant gender effect is observed for English subjects but not for Maths subjects could be the different course study requirements for the calculation of a TES score and an Australian Tertiary Admission Rank (ATAR). In order to receive an ATAR, a student must complete at least 2 units of an approved Board developed course in English. However, it is not compulsory for a student to include a Maths subject in his/her HSC study program to receive an ATAR. Students generally choose an HSC study program that maximises their ATARs.

33 There are 21 Boys and 24 Girls high schools in the NSW DEC system.

to 0.2 standard deviations for senior secondary students. This finding warrants further investigation as to reasons why students appear to achieve more in these schools than in co-educational schools. In the interim, to ensure the fairness of the VA measures, it is proposed to include a single-sex schooling factor in secondary VA models as the establishment of such schools is a system decision, and is out of control of schools or principals.

6 Next steps

The validity of the proposed VA measures depends on our capability to isolate statistically the contribution schools make to students' learning progress, from all other sources of influence. It also depends on the quality of the input measures (including the error associated with measuring contextual factors and student prior and later achievement) and appropriate adjustment given to account for movements of students across schools and systems during a given period of time over which the value added is estimated for a school.

In order to further improve the validity of the VA measures, the following work has been identified and will be undertaken by the Centre for Education Statistics and Evaluation.

Estimating bias arising from movements of students from Year 6 to 7 from government to non-government schools

The proposed VA 5-7 measure for primary schools is based on matching the NAPLAN results of Year 5 students attending a primary school to the NAPLAN results of these students two years later when they were in Year 7. The matching is a 'within system' matching, therefore any students who moved to non-government schools after Year 6 would not be included in the calculation of the VA 5-7 measures for their primary schools. Bias would arise if this group of students are not representative of their respective Year 5 cohorts in their primary schools, on key demographic and achievement indicators that could impact on growth trajectories. Preliminary DEC analysis has not found systemic bias associated with this issue (by comparing the key characteristics of the leavers to those who stayed), but more analysis needs to be undertaken to understand the extent of any potential bias at the school level.

Movements of students across schools

The proposed VA models use only the records of students who stayed in the same school over two consecutive test times. In other words, any students who moved to another government school are not included in the estimation process. This is due to the fact that information on the length of time each student was enrolled in every school he/she moved to over two consecutive test times was not available at the time of the modelling. This means that a 'dosage' approach which appropriately accounts for the extent to which students are exposed to different schools over a target period of time is not able to be tested in the current models. Future work in this area includes system capacity analysis and modelling work involving the use of the 'dosage' approach.

Suitability of teacher assessments at entry to school for K-3 VA measures as baseline indicators

As indicated earlier, exploratory VA K-3 models were developed during the modelling process, using teacher assessments made in the first term of schooling through the Best Start Kindergarten Assessment as baseline indicators. Previous analysis shows a strong relationship between student SES and levels of literacy and numeracy knowledge and skills that each student brings to school as they enter Kindergarten. If this prior ability is not adjusted for, it will lead to bias in VA K-3 estimates and disadvantage schools serving students from low SES backgrounds.

However, as with all other teacher assessments, Best Start teacher assessments may not be consistent and/or comparable which could impact on the validity of the K-3 VA estimates. In this regard, initial analysis using matched records from 2010 Kindergarten to 2013 NAPLAN results shows that, as compared to a 'null' model (i.e., a multilevel model with no explanatory factors included), the inclusion of teacher judgements collected through the Best Start program reduces the student-level unexplained variance by 35 per cent. While this is evidence supporting the

inclusion of Kindergarten teacher assessments in the VA K-3 models, more work examining the quality of the teacher assessments will need to be undertaken. It is also stressed that, as Best Start is designed for a specific purpose, that is to inform the development of quality teaching and learning programs, the use of assessment data collected through that program for the VA analysis purposes will need wide consultation.

Development of non-test outcome based VA estimates

Not all of a school's success can or should be measured through students' achievements in standardised tests. Supplementary measures such as schools' contribution to improve the retention of their students to Year 12 and receiving a HSC; student attendance and behaviour should also be developed and added to the proposed profile of VA measures so that the profile can provide a more balanced and holistic picture of a school's educational excellence.

Ongoing work to identify other, currently unmeasured, contextual factors

Identifying important contextual factors that have an influence on student outcomes but are largely out of schools' control is an important and on-going piece of work to ensure the fairness of the VA measures. As mentioned in Section 4.3, two important contextual factors – EAL/D students with low English language proficiency and students with a confirmed disability are currently not included in the VA models. Further modelling will be undertaken when new data on the language proficiency levels and students with disability become available.

7 Summary

The proposed VA measures for NSW government schools have been designed to be the fairest and most robust measures that are possible given the student assessment and contextual data currently available at the system level.

The key features of the VA models include use of a multilevel modelling approach to more accurately estimate school-level effects; adjusting for those school and student contextual factors that impact on student learning progress; pooling data over more than one year to improve the validity of the measures; and reducing the volatility of VA estimates, for small schools especially, by adjusting estimates in proportion to their reliability.

As for any measure of school performance, VA measures are not perfect nor are they definitive. The proposed VA measures will be most useful when they are used in conjunction with other measures such as absolute performance levels and growth measures to provide a profile of school performance, as well as follow-up in-depth analysis of schools that perform significantly better than others to determine "what works" to improve educational outcomes for students.

8 Appendix

8.1 Appendix 1

Variables tested during the modelling process

Controlling factors	Description
At the student level	
Student prior achievement score	For VA K-3 (exploratory): Students' average assessed level across aspects of literacy/numeracy through the Best Start program in Kindergarten For all other VA measures: Standardised average reading and numeracy scores from previous NAPLAN tests (e.g. Year 3 results in 2010 for the 2010-2012 VA 3-5 measure)
Student's SES measure	Derived from parental highest education, non-school qualification and occupation status sourced from student enrolment forms
Aboriginal and Torres Strait Islander (A&TSI) status	Student's Aboriginal status as collected from the National Schools Statistics collections
Gender	As collected from the National Schools Statistics collections
OC - whether or not a student attended an opportunity class in Year 5 and Year 6	As collected from the National Schools Statistics collections
At the school level	
School size	Average school size over the past three years, derived from data collected for the National Schools Statistics collections
Remoteness of the school	MCEETYA four level remoteness classifications for schools
Proportion A&TSI students in a school	Average proportion of A&TSI students in a school over the past three years, calculated from data collected for the National Schools Statistics collections
School socio-economic status measure	The school Family Occupation and Education Index
Average prior ability level in the school	Prior ability score averaged across matched students in a school
Selective school status	Whether a school is a fully selective school or not (i.e. whether the school streams all its students academically or not)
Single-sex schooling	Whether a school is a boys, girls or a co-educational school

Appendix 2

VA scores with confidence intervals (based on matched records from 2011 to 2013)

Figure 15:

VA 3-5 scores with confidence intervals (based on matched records from 2011 to 2013)

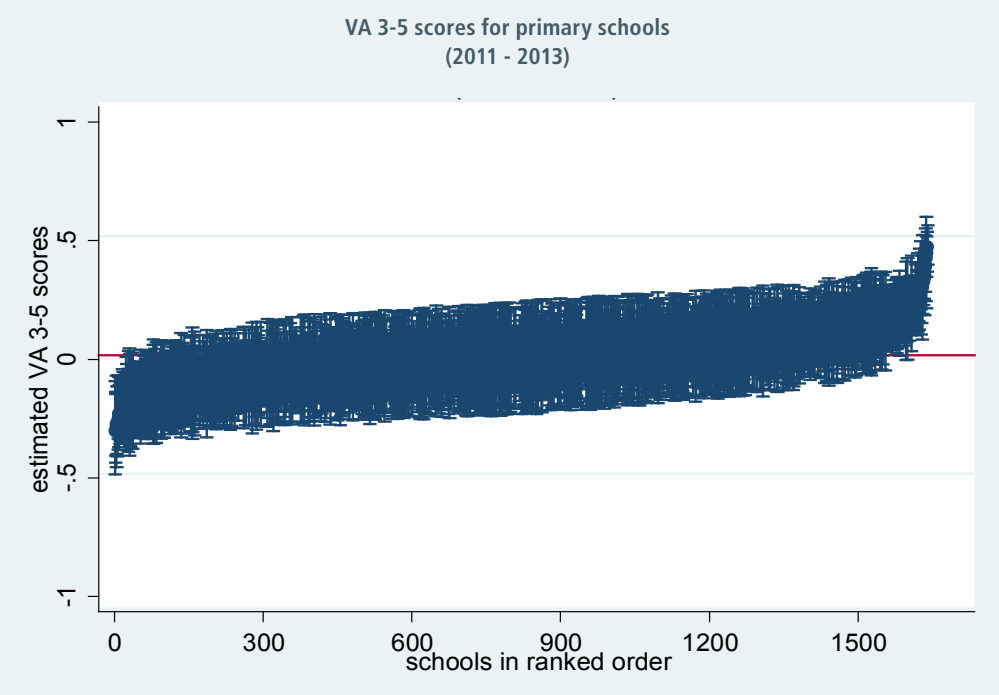
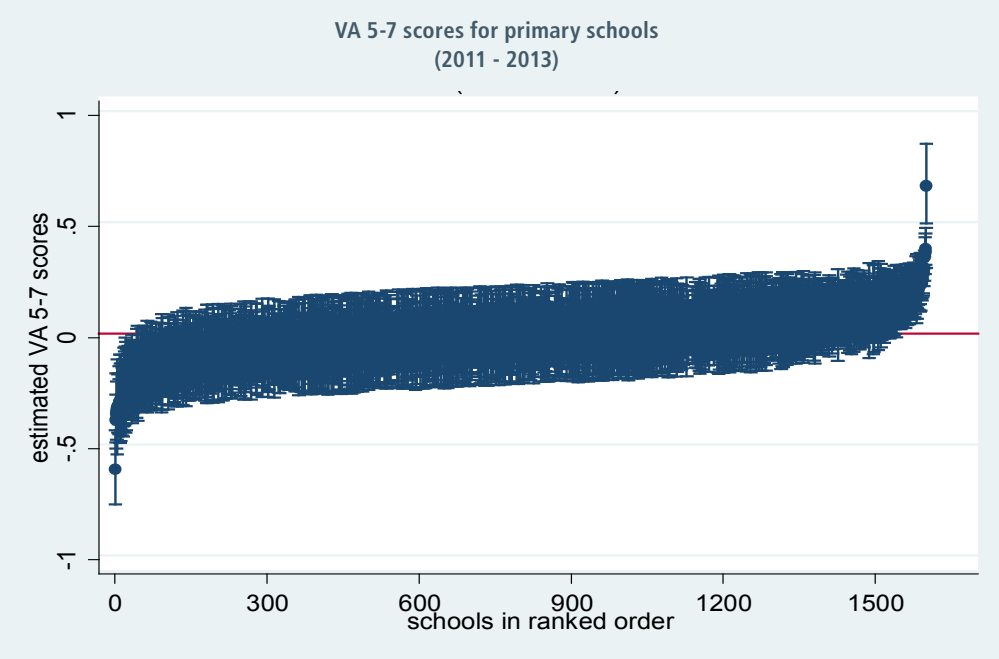


Figure 16:

VA 5-7 scores with confidence intervals (based on matched records from 2011 to 2013)



9 Reference

Aitkin, M., & Longford, N. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A (General)*, 149(1), 1-43.

Baker, D. P., Riordan, C., & Schaub, M. (1995). The effects of sex-grouped schooling on achievement: the role of national context. *Comparative Education Review*, 39(4), 468-82.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioural Statistics*, 29(1), 37-65.

Blank, R. K. (2010). *State growth models for school accountability: progress on developing and reporting measures of student growth*. Washington, D.C: Council of Chief State School Supervisors.

Centre for Greater Philadelphia (CGP) University of Pennsylvania. (2004). *Value-added assessment*. Retrieved March 1, 2014, from http://www.cgp.upenn.edu/ope_value.html

Chetty, R., Friedman, J. N., & Rockoff, J. (2013). *Measuring the impacts of teachers I: evaluating bias in teacher value-added estimates* (Working Paper No. 19423). Cambridge, MA: National Bureau of Economic Research. Retrieved January 15, 2014, from http://www.nber.org/papers/w19423

Cuttance, P. (2001). *The impact of teaching on student learning*. Canberra: Australian College of Educators Yearbook 2000, Australian College of Education.

Deming, D. (2014). *Using school choice lotteries to test measures of school effectiveness* (Working Paper No. 19803). Cambridge, MA: National Bureau of Economic Research. Retrieved May 15, 2014, from http://www.nber.org/papers/w19803

Department of Education and Early Childhood Development (DEECD) Victoria. (2007). *Value-added measures for school improvement*, no.13, Victorian Government, Melbourne.

Department of Education and Early Childhood Development (DEECD) Victoria. (2011). *Measuring performance fairly.* Retrieved 1 April, 2014, from http://www.education.vic.gov.au/Documents/about/department/201112deecdannualreport.pdf

Goldhaber, D., & Hansen M. (2008). Is it just a bad class? *Assessing the stability of measured teacher performance* (Working Paper No. 2008_5). Seattle, WA: Centre on Reinventing Public Education (CRPE).

Griffin, P., Woods, K., & Nguyen, T. (2005). *An environmental scan of tools and strategies that measure progress in school reform*. Report to the Department of Education and Training, Melbourne, Victoria.

Harris, D., & Sass, T. (2006). *Value-Added models and the measurement of teacher quality* (working paper). Tallahassee, FL: Florida State University. Retrieved 15 May 2014, from http://myweb.fsu.edu/tsass/Papers/IES%20Harris%20Sass%20EPF%20Value-added%2014.pdf

Hershberg, T., Adams, Simon, V, & Lea-Kruger, B. (2004). The revelations of value-added, *The School Administrator*, 60(11), 10-14.

Hill, P. W. (1995). Value-added measures of achievement. *Incorporated Association of Registered Teachers of Victoria* (IARTV) Seminar Series, no. 44.

Hong Kong Education Bureau. (2012). *Schools Value-added Information System Technical Manual (SVAIS).* Retrieved January 15, 2014, from https://svais.edb.gov.hk/Content/docs/en/TechnicalManual.pdf.

Isenberg, E., & Hock, H. (2011). *Design of value-added models for IMPACT and TEAM in DC public schools*, 2010-2011 school year. Washington D.C: Mathematical Policy research.

Johnson, M., Lipscomb, S., Gill, B., Booker, K., & Bruch, J. (2012). *Value-added models for the Pittsburgh public schools*. Cambridge, MA: Mathematical Policy Research.

Kane, T., & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, 16(4), 91-114.

Kane, T., & Staiger, D. (2002). Volatility in school test scores: implications for test-based accountability systems. In D. Ravitch (Ed.), *Brookings papers on education policy: 2002* (pp. 235-283). Washington, D.C: Brookings Institution Press.

Kane, T. J., & Staiger., D. O. (2008). *Estimating teacher impacts on student achievement: an experimental evaluation* (Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research. Retrieved 15 May 2014, from http://www.nber.org/papers/w14607

Kane, T.J., McCaffrey, D.F., Miller, T., & Staiger, D.O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment* (Measuring Teacher Effectiveness (MET) Project Research Paper). Seattle, WA: Bill & Melinda Gates Foundation. Retrieved 15 May, 2014, from http://www.metproject.org/downloads/MET_Validating_Using_Random_Assignment_Research_Paper.pdf

Leckie, G. (2013). England's multilevel model based value-added school league tables: measuring and communicating statistical uncertainty to parents. *Bulletin of the International Statistical Institute*, 68, 1-6.

Loeb, S., & Candelaria, C. (2013). *How stable are value-added estimates across years, subjects and student groups?* (Carnegie Knowledge Network, Knowledge Brief 3). Stanford, CA: Carnegie Foundation for the Advancement of Teaching.

Massachusetts Department of Elementary and Secondary Education. (2011). *Massachusetts Comprehensive Assessment System (MCAS) student growth percentiles: interpretive guide.* Retrieved January 15, 2014, from http://www.doe.mass.edu/mcas/growth/interpretiveguide.doc

McCaffrey, D., Sass, T., Lockwood, J.R., & Mihaly, K. (2009). The intertemporal stability of teacher effects. *Education Finance and Policy*, 4(4), 572-606.

Muijs, D., & Reynolds, D. (2001). *Effective teaching: evidence and practice*. London: Paul Chapman Publishing.

Nuttall, D.L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. In B. Creemers & J. Scheerens, J (Eds.). *Developments in school effectiveness research, special Issue of International Journal of Educational Research*, 13, (7), 769-776.

Organisation for Economic Co-operation and Development (OECD). (2006). *Women in scientific careers: unleashing the potential*, Paris: OECD Publishing.

Organisation for Economic Co-operation and Development (OECD). (2008). *Measuring improvements in learning outcomes: best practices to assess the value-added of schools*, Paris: OECD Publishing.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2005). *Multilevel analysis with MLwiN software: a user's guide to MLwiN(2)*. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Raudenbush, S., & Bryk, S. (2002). Hierarchical linear models, applications and data analysis methods (2nd ed.). London: SAGE Publications.

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Journal of Economics*, 125(1), 175-214.

Rowan, B., Correnti, R., & Miller, R. J. (2002). *What large-scale, survey research tells us about teacher effects on student achievement: insights from the prospects study of elementary schools* (CPRE Research Report Series RR-051). Philadelphia, PA: Consortium for Policy Research in Education.

Rowe, K. J. (2004, August). *The importance of teaching: ensuring the quality of teaching and learning provision by building teacher capacities that maximise students' positive experiences and outcomes of schooling - implications of findings from emerging international and Australian evidence-based research.* Supporting paper delivered at the Making Schools Better Conference: A Summit Conference on the Performance Management and Funding of Australian Schools, Melbourne, Victoria.

Rowley, G. (2006). *Value-added measures in education and training*. Melbourne: report to the Department of Education and Training, Victoria.

Salomone, R. (2003). *Same, different, equal: rethinking single-sex schooling*. New Haven: Yale University Press.

Sammons, P., Nuttall, D., & Cuttance, P. (1993). Differential school effectiveness: results from a re-analysis of the Inner London Education Authority's (ILEA) junior school project data. *British Educational Research Journal,* 19(4) 381-405.

Sammons, P., Thomas, S., Mortimore, P., Owen, C., & Pennell., H. (1994). *Assessing school effectiveness: developing measures to put school performance in context*. London: Office for Standards in Education (OFSTED).

Sanders, W.L. (2000). Value-added assessment from student achievement data: opportunities and hurdles. *Journal of Personnel Evaluation in Education*, 14(4), 329-339.

Schochet, P., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington D.C: National Centre for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Sikora, J. (2013). *Single-sex schools and science engagement*. Adelaide: National Centre for Vocational Education Research (NCVER) Occasional Paper.

Snijders, T., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22(3), 342-363.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis*. London: Sage Publications.

Streitmatter, J. (2002). Perceptions of a single-sex class experience: females and males see it differently. In A. Datnow & L. Hubbard (Eds.), *Gender in policy and practice: perspectives on single-sex and coeducational schooling* (pp.212-226). New York: Routledge Falmer.

Webster, W. J. (2005). The Dallas School-Level Accountability Model: The Marriage of Status and Value-Added Approaches. In R. Lissitz (Ed.), *Value added models in education: Theory and Applications*. Maple Grove, MN: JAM Press.

Further Information
For further information relating to the report please contact:

Dr. Lucy Lu
Statistics Unit
Centre for Education Statistics and Evaluation
NSW Department of Education and Communities
T 02 9561 8691
F 02 9561 8055
www.dec.nsw.gov.au/cese

NSW GOVERNMENT Education & Communities