THE UNIVERSITY OF MELBOURNE

FACULTY OF BUSINESS & ECONOMICS

Melbourne Institute Working Paper Series

Working Paper No. 7/13

Comparing Least-Squares Value-Added Analysis and Student Growth Percentile Analysis for Evaluating Student Progress and Estimating School Effects

Brendan Houng and Moshe Justman

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Comparing Least-Squares Value-Added Analysis and Student Growth Percentile Analysis for Evaluating Student Progress and Estimating School Effects*

Brendan Houng[†] and Moshe Justman[‡]

[†] Melbourne Institute of Applied Economic and Social Research, The University of Melbourne

[‡] Melbourne Institute of Applied Economic and Social Research, The University of Melbourne; and Department of Economics, Ben Gurion University, Israel

Melbourne Institute Working Paper No. 7/13

ISSN 1328-4991 (Print)

ISSN 1447-5863 (Online)

ISBN 978-0-7340-4296-5

March 2013

* This research uses data from the Victorian Department of Education and Early Childhood Development (DEECD) under the Research and Evaluation Partnerships Agreement with the Melbourne Institute of Applied Economic and Social Research. Thanks to Mike Helal for his part in setting up the NAPLAN data and to Damian Betebenner for helping us adapt his software to Australian data. Thanks to Henry Braun, Cain Polidano, Chris Ryan, and seminar participants at the Melbourne Institute and the DEECD for their helpful comments and suggestions. The views expressed in this paper are those of the authors alone and do not represent those of DEECD. Corresponding author, <justman@bgu.ac.il>.

Melbourne Institute of Applied Economic and Social Research
The University of Melbourne
Victoria 3010 Australia
***Telephone* (03) 8344 2100**
***Fax* (03) 8344 2111**
***Email* melb-inst@unimelb.edu.au**
***WWW Address* http://www.melbourneinstitute.com**

Abstract

This paper compares two functionally different approaches to analyzing standardized test data: least-squares based value-added analysis, geared principally to supporting teacher and school accountability; and Betebenner's (2009) student growth percentiles, which focuses primarily on tracking individual student progress in a normative context and projecting probable trajectories of future performance. Applying the two methods to Australian standardized numeracy and reading test scores (NAPLAN) in grades 3 to 5 and 7 to 9, we find that although they are used differently, the two methods share key structural elements, and produce similar quantitative indicators of both individual student progress and estimated school effects.

JEL classification: I21, I28
Keywords: Value-added analysis, student growth percentiles, NAPLAN

1. Introduction

Standardized testing in elementary and secondary schools creates the possibility of tracking students' academic achievement as they advance through school, comparing it to expected achievement, and using the comparison in forming teacher and school evaluations. Initial efforts in this vein, notably the extensive testing mandated by No Child Left Behind in the United States, set expectations of student achievement in reference to general, grade-specific levels of proficiency without regard to students' individual circumstances. As students' starting points are strongly affected by the home environment, this placed teachers and schools serving weaker populations at an unfair disadvantage (Goldstein and Spiegelhalter, 1996).

Recent approaches to analyzing student performance on standardized tests address this shortcoming by forming *student-specific* expectations of current achievement based on individual past performance, to which they compare actual achievement. School or teacher effects can then be estimated by aggregating the disparities between actual and expected achievement of students in a school, or assigned to a teacher, possibly taking into account other relevant variables describing the student, or the circumstances of specific classrooms or schools.

This paper compares two leading methods of implementing this approach. The first, more widely used of these is value-added analysis based on least-squares regression (LS-VAA). It regresses current scores on prior scores to form conditional expectations of student outcomes; calculates the difference between actual and predicted scores as its normative measure of student progress; and uses fixed effects or random effects (hierarchical linear models) to identify the contribution of individual schools or teachers, which can be roughly characterized as the average residual between actual and expected scores of the students in a

school or the students assigned to a particular teacher.[1] A second more recent approach, which is rapidly gaining popularity, is Betebenner's (2009) student growth percentile analysis (B-SGP). It estimates a set of one hundred quantile regressions—one for each percentile from .005 to .995—conditioned on prior scores. This produces for each student an individual "growth chart," which allows a conditional percentile ranking to be assigned to her actual current score; this is the student's growth percentile.[2] Schools can then be ranked by the median percentile rank of students in the school, or by other appropriate metrics such as the average probability of a student in the school achieving a required level of proficiency.

We compare these two approaches by applying them to standardized test data from Australia's National Assessment Program - Literacy and Numeracy (NAPLAN), focusing on two knowledge domains, numeracy and reading, and on two cohorts of students in Victoria state schools: those studying in grades 3 and 7 in 2008, and progressing respectively to grades 7 and 9 in 2010. We compare the two methods in terms of both their individual student evaluations and the school effects they produce.[3] Specifically, we compare the accuracy of their projections, conditioned on prior scores in all five NAPLAN domains—numeracy, reading, writing, spelling and grammar—and the similarity of their individual student evaluations; and we consider how similar are the school effects they produce, in general and

[1] See Harris (2011), Dearden et al. (2011), National Research Council (2010), Ray et al. (2009), McCaffrey et al. (2009), OECD (2008), Braun (2005), McCaffrey et al. (2004).

[2] See Betebenner (2011) for further details. The standard application estimates cubic B-splines with four internal knots and adjusts the result to eliminate crossing of quantile contours. As Castellano and Ho (2011) note, the term "student growth percentile" is misleading; they describe them as "conditional status percentile ranks".

[3] The NAPLAN database does not support analysis of teacher effects at this point.

in identifying exceptionally high-performing and under-performing schools, and compare their sensitivity to outlying student observations.

Our results indicate that they perform similarly in all these dimensions, confirming, complementing and extending the previous work of Briggs and Betebenner (2009), which reports correlations between the two types of school effects in a paper focused on comparing their sensitivity to monotonic transformations of test scores; Wright's (2010) comparison of *teacher* effects derived from these two different approaches; and Castellano and Ho's (2012) careful comparison of student-level percentile ranks of residuals (PRRs) derived from OLS regressions to student growth. Where our analyses coincide they reach similar conclusions.

That LS-VAA and B-SGP should produce very similar results is not immediately apparent. There are clear methodological differences between them. As Reardon and Raudenbush (2008) point out, LS-VAA assumes that tests are graded on an additive interval scale—this is implicit in averaging residuals to derive school effects—where B-SGP makes no such assumption.[4] B-SGP is by definition invariant to positive monotonic transformations of current test scores, the left-had variable in the quantile regressions it estimates, where least squares regression analysis is sensitive to such transformation.[5] In addition, the linearity and homoscedasticity assumed in regression analysis renders it more accurate than B-SGP where

[4] See also Bond and Lang (2012), Mariano et al. (2010), Ballou (2009), Briggs, et al. (2008) and Young (2006) on the assumption of a vertical scale.

[5] However, Briggs and Betebenner's (2009) analysis of Colorado reading test scores finds that when scale transformations of the type "one would not consider surprising" are applied to the right-hand variables, correlations between LS-VAA school effects derived from test scores before and after these transformations range between .87 to 1.00. When an exponential transformation is applied correlations fall sharply to a range between .30 and .66.

these assumptions hold but less accurate where they do not hold—often at the extremes of the test-score distribution, where scores are shaped by floor and ceiling effects. And there is reason to expect median-based school effects derived from B-SGP to be less sensitive to outlying observations than mean-based school effects derived from LS-VAA.

These differences notwithstanding, our empirical analysis indicates that the two methods produce very similar student evaluations as well as very similar school effects. The correlations between OLS student-level residuals and B-SGP student percentiles within cohort-domain pairs are all in the vicinity of .95; and the mean squared deviation of conditional means produced by OLS regressions from actual scores is nearly identical to the mean squared deviation of B-SGP conditional medians from actual scores. These findings are consistent with Castellano and Ho's (2012) analysis of data from two states in the United States, which finds strong positive correlations between students' percentile ranks of residuals (PRRs) from OLS regressions and B-SGP student percentiles.

School effects derived from the two methods are also very similar, both overall and in identifying the low and high ends of the distribution. Correlations of the two sets of school effects across schools within cohort-domains range between 0.89 and 0.95. These findings are consistent with Briggs and Betebenner (2009) who estimate LS-VAA and B-SGP school effects from Colorado reading test data, conditioned on one, two and three years of prior reading scores, and find correlations of .72, .84 and .91 between them.[6] They are also generally consistent with Wright's (2010) comparison of teacher effects on math scores estimated from LS-VAA and B-SGP using multiple years of prior data, and finds a correlation of .90 between them.

[6] They use multiple prior years of reading scores. Our analysis is conditioned on multiple tests scores—in numeracy, reading, writing, grammar and spelling—from a single prior year.

Next, applying each method to estimate a 95% confidence interval for each school effect, and using this to classify schools as significantly better or worse than the median school or not significantly different from it, we find that between 86 and 91 percent of schools in each cohort-domain are similarly classified by both methods, and no school classified as significantly above the median by one method is identified as significantly below by the other. Moreover, the two methods are also similar at the extremes of the distribution with extensive overlap between the ten highest-ranked schools identified by LS-VAA and by B-SGP; and between the two lists of the ten lowest ranked schools each identifies. Finally, we found that B-SGP school effects are only slightly less sensitive to outlying student observations: correlations of school effects with and without outliers range from .94 to .96 for LS-VAA and from .97 to .98 for B-SGP.

There are well-recognized differences between LS-VAA and B-SGP in the way they are used in practice. LS-VAA specializes in teacher and school accountability, investing considerable effort in methodological development; B-SGP focuses on presenting teachers and parents with a clear, informative picture of student progress, and a grounded projection of likely future achievement. Consequently, education systems that value standardized tests for their contribution to teacher and school accountability favor LS-VAA, while those that view standardized testing primarily as a tool for educators and parents to improve teaching and learning, tend to favor B-SGP analysis. Our findings indicate that their statistical similarity is such that LS-VAA could be applied to tracking and projecting student growth and B-SGP to school accountability with little change in the results—though we cannot say anything about applying B-SGP to teacher accountability.[7]

[7] See National Research Council (2010, 2011) on concerns regarding the use of standardized testing for teacher accountability some of which applies, in lesser measure, also to school

In the remainder of the paper, Section 2 presents the structure of the data and some descriptive statistics; Section 3 compares the accuracy of individual predictions; Section 4 describes the two sets of school effects and compares them to each other, with special emphasis on high-performing and low-performing schools, and sensitivity to outliers; and Section 5 concludes.

2. Data

Since 2008, student achievement in Australia's schools has been monitored by the National Assessment Program – Literacy and Numeracy (NAPLAN), comprising standardized tests administered annually to all students in grades 3, 5, 7 and 9 in five knowledge domains: numeracy, reading, grammar, spelling and writing, taken over a single week in mid-May (the school year begins in early February). NAPLAN scores are scaled on an integrated vertical scale in such a way that test results can be compared across grade levels, and calibrated to a constant level of difficulty so that scores can be compared over time. Tests on reading, spelling, grammar and numeracy are in the format of multiple choice questions while writing tasks are marked using structured procedures to maintain consistency. We focus our analysis on student progress in Victoria state schools in numeracy and reading, for the cohorts studying in grades 3 and 7 in 2008 and progressing respectively to grades 5 and 9 in 2010, using spelling, grammar and writing scores only as right-hand variables. See Helal (2012) for further details on NAPLAN.

---

accountability. These include instability across time; selection bias; the difficulty in separating teacher effects from school effects; the limited scope of application of these tests to selected grades and subjects; incentives for narrowing the curriculum; and the difficulty parents, educators and the public will have understanding how they work leading to misuse.

Our data comes from the student performance database of the Department of Education and Early Childhood Development (DEECD) from which we constructed matched cohorts for 2008-2010, matching 2008 grade 3 and grade 7 student outcomes to 2010 grade 5 and grade 9 outcomes, respectively. Table 1 describes the construction of the estimation sample used in our empirical analysis from the full student population. In Victoria, 67 percent of primary school students and 57 percent of secondary school students attend state schools. The majority of state schools in Victoria are stand-alone primary or secondary schools, with primary schools offering education from prep to grade 6 and secondary schools covering grade 7 to grade 12.

In Table 1, column (a) indicates that in 2008 there were approximately 44,500 students in Victoria state schools in grade 3 and 38,700 in grade 7, reflecting substantial movement from the state sector to the private sector (independent or Catholic) in the transition from primary to secondary school. We drop students who exit the state system or transfer to another school within the state system between grade 3 and 5 or between grade 7 and 9 (column b); and students for whom we don't have a full set of test scores in both grades (column c). The remaining students—those who remained in the same state school in the relevant time period and for whom we have valid test scores in all five domains for both grades—are tallied in column (d). They are the students for whom we predict performance in the later grade based on test scores in the earlier grade, and we refer to them as our student prediction sample; we use it to compare the accuracy of predictions and to identify outliers.

In estimating school effects, we omit schools with fewer than 20 students in the cohort in the student prediction sample (column e), leaving us with 649 schools for the primary school cohort and 260 schools in our secondary school cohort. The final numbers in column (f) are our school-effects sample which we use to re-estimate the regression equations with school

fixed effects and to calculate median SGPs for each school. We find attrition rates of just over 30 percent in primary school and about 40 percent in secondary school from the full cohort to the prediction sample; and a further decline of 12 percent in the transition from the prediction sample to the school based sample in primary school, and 1 percent in secondary school.

Table 1 Numbers of students in the population and in the estimation samples

	Cohort size	Exits and transfers	Incomplete set of test scores	Prediction sample	% of cohort	In schools with < 20 students	School-effects sample	% of cohort
	(a)	(b)	(c)	(d)=(a)-(b)-(c)	(d)/(a)	(e)	(f)=(d)-(e)	(f)/(a)
Grade 3 - 5	44,578	8,716	5,205	30,657	69%	4,288	26,369	59%
Grade 7 - 9	38,742	7,801	7,505	23,436	60%	420	23,016	59%

(b) Students who exit the state school system or transfer to another state school in 2008-10
(c) Students who have less than a full set of valid NAPLAN scores for all domains in both grades
(f) Students enrolled in a school with fewer than 20 students in the prediction sample in their cohort

The rates of attrition presented in Table 1 are substantial and may well be non-random but as Table 2 shows, the full cohort, the prediction sample and the school sample all have similar observable characteristics, as measured in 2008. Moreover, as Table 3 shows, the means and standard deviations of the test scores in each grade and domain are roughly comparable across the three groups. Differences in the averages between the full cohort averages and the prediction sample range from between 5 and 7 points in grade 3 and between 2 and 4 points in grade 9. Differences in the means between the prediction sample and the school-effects sample are never more than 3 points. Standard deviations are also slightly larger for the full cohort than for the prediction sample and virtually identical for the prediction and school-effects sample. Altogether we interpret this as indicating that despite the substantial attrition noted above, the characteristics of the prediction and school-effects samples are similar to those of the original population.

Table 2 Student demographic characteristics

Grade 3, 2008	All students	Prediction sample	School-effects sample
Male	51.5%	50.4%	50.3%
Language background other than English	24.1%	23.5%	24.8%
Aboriginal or Torres Strait Islander origin	1.5%	1.1%	0.9%
Both parents unemployed in last 12 months	6.1%	4.9%	4.4%
Mother or father a manager or senior manager	40.5%	43.8%	45.8%
Mother or father has tertiary education	25.0%	27.1%	29.1%
Neither parent has more than a grade 10 education	5.5%	4.9%	4.4%
Grade 7, 2008			
Male	52.3%	51.7%	51.7%
Language background other than English	25.5%	26.7%	27.0%
Aboriginal or Torres Strait Islander origin	1.7%	0.9%	0.9%
Both parents unemployed in last 12 months	6.1%	5.4%	5.4%
Mother or father a manager or senior manager	32.6%	35.1%	35.1%
Mother or father has tertiary education	17.8%	19.2%	19.4%
Neither parent has more than a grade 10 education	7.4%	6.9%	6.9%

Table 3 Mean scores in the full population and in the prediction sample
(*standard deviations in parentheses*)

	Numeracy	*Reading*	*Grammar*	*Spelling*	*Writing*
Grade 3, 2008					
Full cohort	416.14	415.82	423.40	411.18	421.25
	(72.53)	(82.24)	(87.69)	(75.15)	(68.64)
Prediction sample[a]	421.20	421.71	430.07	417.02	426.34
	(71.07)	(79.61)	(84.21)	(72.17)	(66.21)
School-effects sample[b]	422.37	423.68	432.16	420.26	428.56
	(71.04)	(79.28)	(83.72)	(71.67)	(66.08)
Grade 5, 2010					
Full cohort	501.78	497.91	506.98	490.97	492.50
	(72.41)	(77.61)	(83.52)	(67.62)	(67.46)
Prediction sample[a]	507.19	503.61	513.08	495.21	497.27
	(71.51)	(76.55)	(81.23)	(65.96)	(66.02)
School-effects sample[b]	508.52	504.93	514.86	497.31	499.69
	(71.85)	(76.39)	(80.71)	(65.42)	(66.13)
Grade 7, 2008					
Full cohort	541.21	531.11	524.29	531.02	534.16
	(71.10)	(67.51)	(78.36)	(70.52)	(83.18)
Prediction sample[a]	546.51	535.90	530.14	535.93	540.83
	(68.52)	(65.73)	(75.31)	(67.32)	(78.51)
School-effects sample[b]	546.69	535.98	530.35	536.32	541.01
	(68.59)	(65.65)	(75.33)	(67.31)	(78.53)
Grade 9, 2010					
Full cohort	583.42	570.35	572.15	570.36	563.55
	(71.60)	(66.26)	(77.27)	(75.97)	(92.40)
Prediction sample[a]	585.61	574.25	575.27	572.52	568.39
	(66.28)	(63.82)	(72.20)	(72.12)	(86.74)
School-effects sample[b]	585.81	574.35	575.51	572.85	568.56
	(66.35)	(63.82)	(72.28)	(72.08)	(86.72)

[a] Students who have a valid score for all domains and were in the same public school in 2008 and 2010.
[b] Students in the prediction sample attending a school with at least twenty students in the year.

3. Prediction and assessment of individual performance

Both LS-VAA and B-SGP begin by setting individual standards for each student's current performance based on her prior scores; then compare students' actual performance to their individual standards; and finally calculate a summary statistic of these divergences of actual from standard performance for each school. In LS-VAA these individual standards are the conditional means, or predicted scores, derived by OLS regression of current scores on prior

scores. In B-SGP analysis, the conditional medians are the counterpart of the conditional means—they are the norm against which current performance is measured.

Throughout our analysis we omit student demographic and socio-economic covariates, such as gender or parents' education, from the analysis. There is a strong *a priori* argument for omitting them, as this would imply having different expectations from students *with the same prior scores*, depending on their gender or socio-economic background. In this we follow leading applications of LS-VAA (McCaffrey et al., 2004; Sanders and Horn, 2005; Ballou et al., 2004) and Betebenner's (2009) SGP analysis. They find that empirically the loss of precision is minimal, as do we. Of course, this does not preclude incorporating background variables in interpreting student outcomes or school rankings, as Ehlert et al. (2012) strongly advocate.

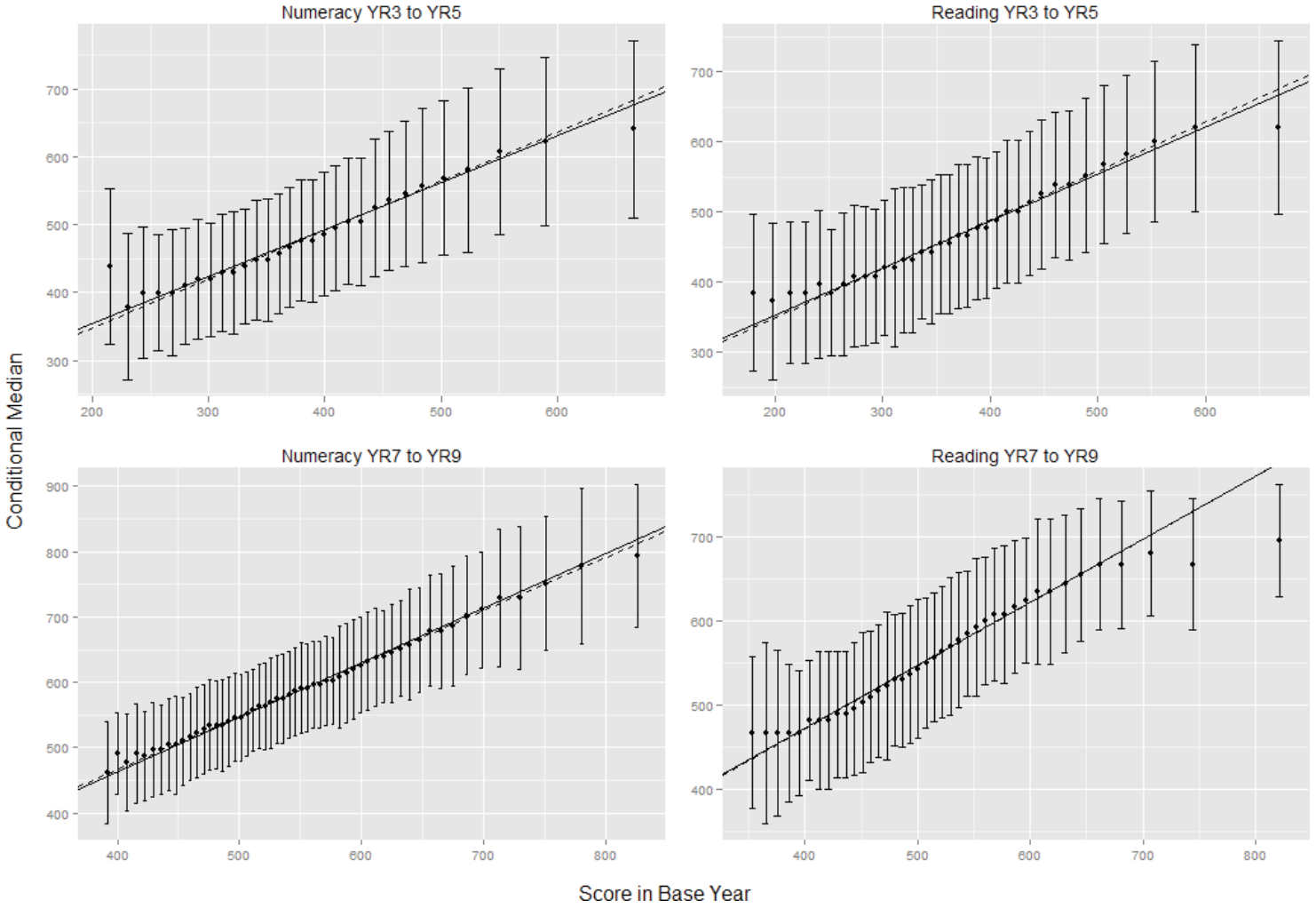
Before presenting our full comparison of LS-VAA and B-SGP we offer a graphic comparison between two slimmed-down versions of these methods that condition expected current scores only on a single prior score in the same domain; this allows a two-dimensional graphic representation. The results of the OLS estimations are presented in Table 4, in the column marked simple regression, for each of the four cohort-domain pairs. The equation is highly significant and explains almost as much of the variance as the multiple regressions reported in the adjacent columns.[8] The coefficients of prior scores in the same domain are estimated precisely and are all significantly less than one. Predictions are more accurate for secondary than for elementary school and more accurate for numeracy than for reading.

[8] Adding gender and SES indicators increased the R^2 by no more than .014 for any cohort-domain, confirming previous findings that they have little effect after controlling for prior scores.

Table 4 Numeracy and reading regressions, *standard errors in parentheses*

	Grade 3 to 5		Grade 7 to 9	
Numeracy skills	simple	mutliple	Simple	multiple
Numeracy	0.722	0.563	0.812	0.739
	(0.004)	(0.006)	(0.003)	(0.005)
Reading		0.068		0.044
		(0.006)		(0.005)
Grammar		0.043		0.022
		(0.005)		(0.005)
Spelling		0.113		0.036
		(0.006)		(0.005)
Writing		0.027		0.025
		(0.005)		(0.004)
Constant	203.128	164.399	141.758	113.901
	(1.710)	(2.058)	(1.888)	(2.230)
R^2	*0.515*	*0.542*	*0.705*	*0.712*
Reading skills				
Numeracy		0.217		0.165
		(0.006)		(0.005)
Reading	0.696	0.441	0.751	0.502
	(0.004)	(0.006)	(0.004)	(0.006)
Grammar		0.119		0.109
		(0.005)		(0.005)
Spelling		0.005		0.002
		(0.006)		(0.005)
Writing		0.072		0.079
		(0.006)		(0.004)
Constant	210.258	142.352	171.670	113.695
	(1.623)	(2.130)	(2.173)	(2.380)
R^2	*0.525*	*0.572*	*0.598*	*0.647*

Figure 1. Conditional medians with 95% confidence intervals, trendlines and regression lines



The conditional medians for each prior score are calculated directly and plotted in Figure 1, along with 95 percent confidence intervals and a fitted slope denoted by the solid line, to which we added the regression slope, represented by a broken line. NAPLAN derives test scores by estimating a Rasch model such that the number of different test scores equals the number of test items. Consequently, each score has a frequency of many hundreds, except at the extremes of the distribution. In each of four graphs the trendline fitted to the conditional medians and the OLS regression line almost coincide. The conditional medians follow the same linear trends as the conditional means except at the ends of the distribution where differences among these prior scores seem to have little or no predictive power, presumably reflecting floor and ceiling effects.[9] The confidence intervals are similar in size within each cohort-domain, except for moderate increases in variance at the extremes.

Next we examine, graphically and numerically, the joint distribution of indicators of individual progress that the two methods produce, conditioned all prior scores. Figure 2 plots the standardized residuals from the multivariate regressions on the y-axis against the individual growth percentiles produced by the B-SGP analysis on the x-axis. In Table 5, each entry represents the share of students similarly classified by both methods, with the rightmost entry giving the total share of students classified similarly by both methods; it ranges between 87.2 and 89.6 percent. Less than 1 percent of students in any cohort-domain were classified more than one category apart by the two methods.

[9] Some of the high-end scores seem to be inconsistent with NAPLAN guidelines. In grade 7 numeracy we found a small number of very low frequency scores in the middle of the distribution which we combined, in Figure 1, with adjacent high-frequency prior scores.

Figure 2. Individual residuals plotted against student growth percentiles

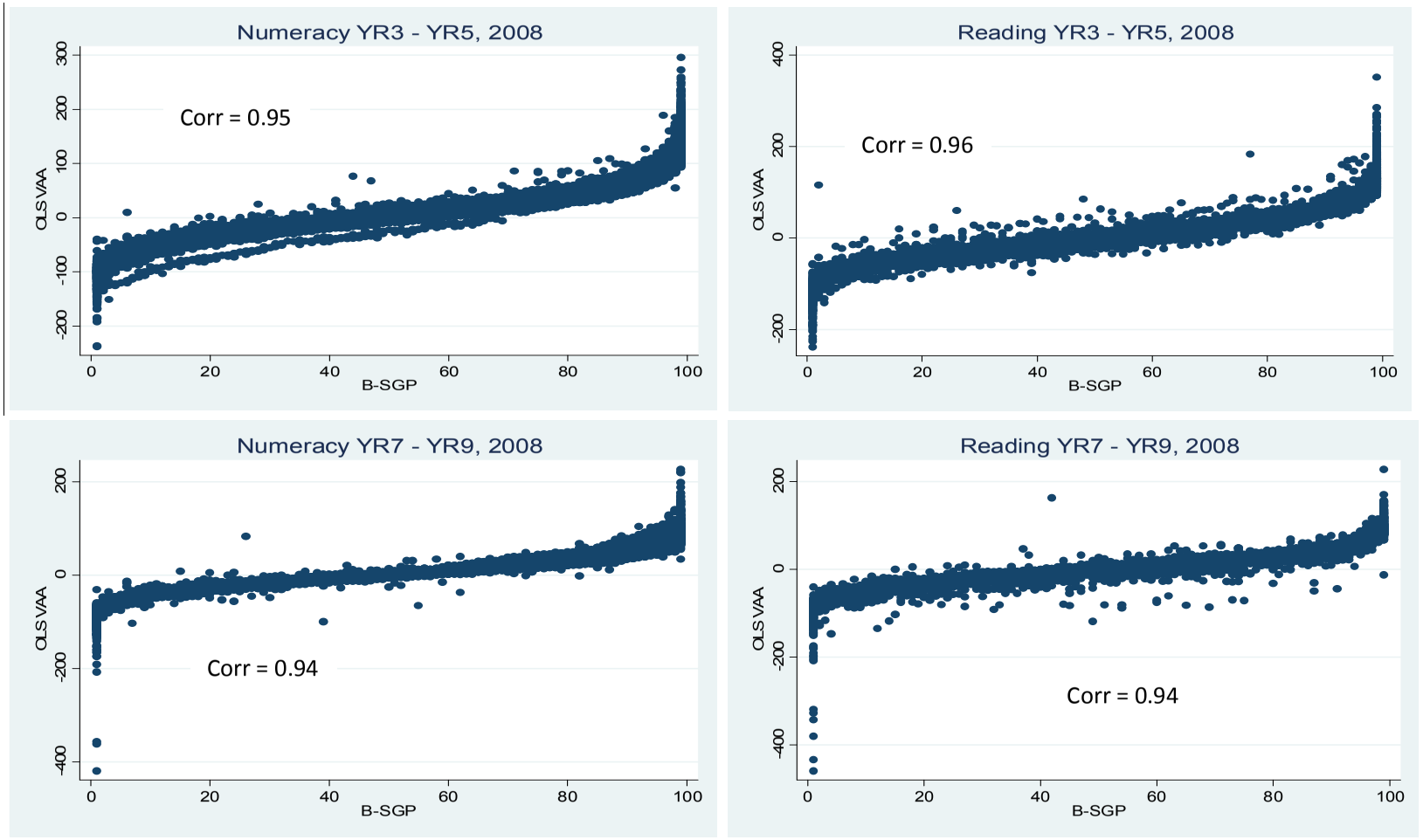


Table 5. Joint frequency of student percentile ranks for each grade-domain

	0%-2.5%	2.5%-10%	10%-25%	25%-75%	75%-90%	90%-97.5%	97.5%-100%	*sum*
Grade 3 to 5 Numeracy	1.93	6.18	12.69	46.98	12.03	5.59	1.97	87.4
Grade 3 to 5 Reading	2.10	6.61	12.71	47.44	12.80	5.96	1.99	89.6
Grade 7 to 9 Numeracy	2.17	6.53	12.85	47.22	12.29	5.35	1.83	88.2
Grade 7 to 9 Reading	2.06	6.18	11.98	46.44	12.37	5.97	2.17	87.2

Finally, Table 6 compares the accuracy of the two methods, the extent to which actual scores diverge from the conditional means and medians produced by the two methods, using the R^2 statistic, defined conventionally for OLS-VAA, and correspondingly for B-SGP as:

$$Q_{gd} = 1 - \sum_i (s_{igd} - \hat{s}_{igd})^2 / \sum_i (s_{igd} - \bar{s}_{gd})^2$$

where s_{igd} is the actual score of student i in grade g and domain d; $\hat{s}_{igd}$ is her conditional median, based on prior scores; and $\bar{s}_{gd}$ is the average score in grade g and domain d. The results presented in Table 6 indicate that these measures are nearly identical across methods.

Table 6. Comparison of R^2 values between actual and expected scores

	Grade 3 to 5		Grade 7 to 9	
	OLS-VAA	B-SGP	OLS-VAA	B-SGP
Numeracy	0.54	0.55	0.71	0.71
Reading	0.57	0.58	0.65	0.66

Expected scores are conditional means for OLS analysis and conditional medians for SGP analysis.

4 School effects

We now turn to compare school-level indicators derived from LS-VAA and B-SGP, focusing our analysis on schools with at least 20 students in each cohort. We estimate OLS-based school effects by including school fixed effects in OLS regressions of students' current scores on past scores. Differences between school effects are essentially differences in the average residuals (actual minus predicted scores) of the students in each school.[10] We estimate SGP-based school effects as the median growth percentile of all students in the school.

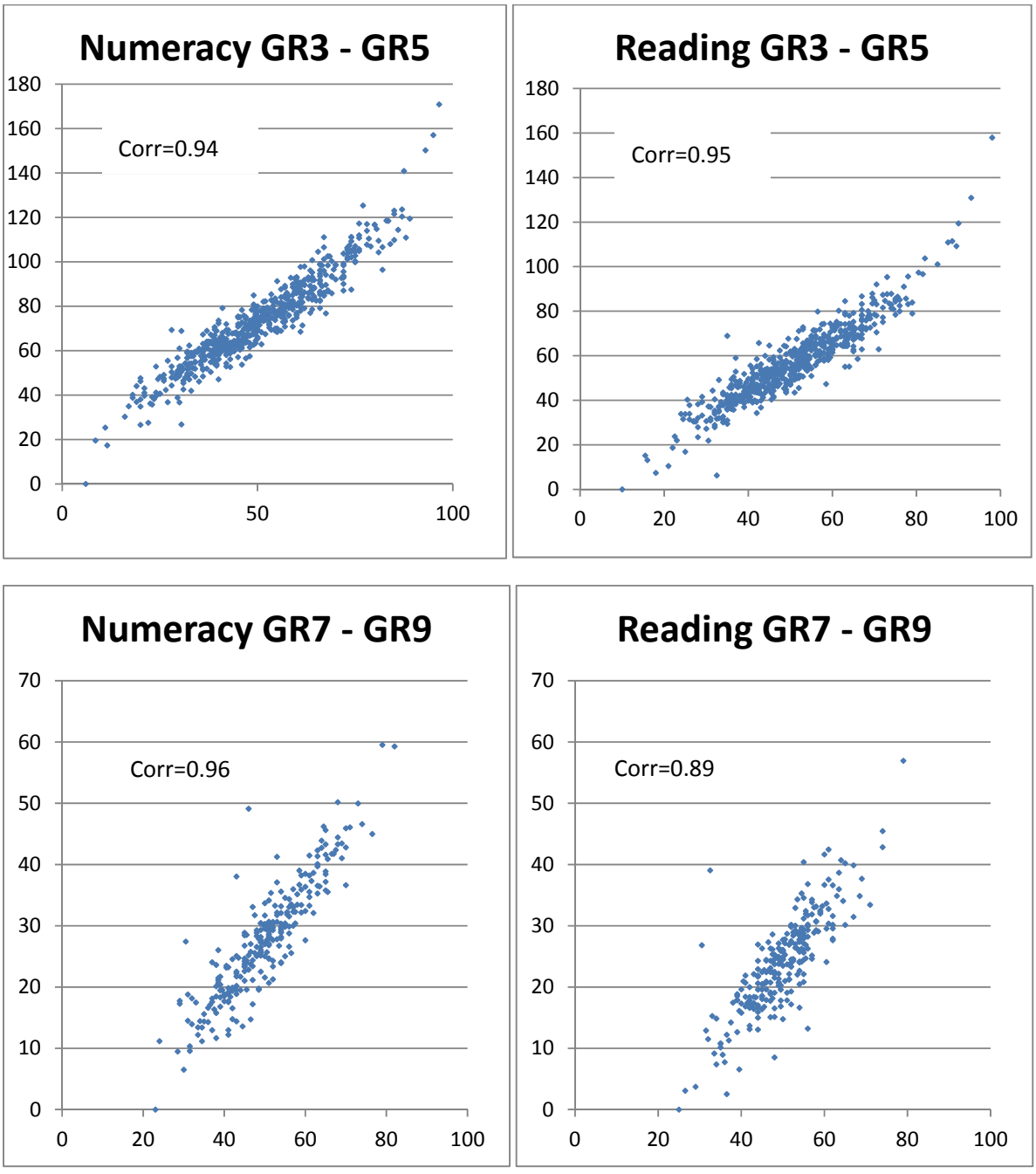
First we consider the general similarity of the two sets of school effects graphically and by calculating the correlations between them. Then we ask, to what extent are the two methods similar in identifying the highest and lowest ranking schools? We answer this on two levels. We first identify schools that place significantly above or below the average or median school by each method and tabulate the joint frequencies of schools significantly above or below the median outcome. Then we identify (without naming) the top and bottom ten schools in each category by each method, and compare the two rankings. Finally, we consider the effect of outliers on the stability of school effects by identifying, for each method, students whose individual progress indicators fall in the top or bottom 2.5%, recalculating the two sets of school effects excluding these students, and comparing the results with the original school effects derived from the full population.

---

[10] The omitted school, which serves as our reference point, is the school with the median effect. We also estimated a random effects model. While the random effects model failed the Hausman test we found the random and fixed effects to be almost perfectly correlated.

Figure 4 presents scatterplots and correlations of the two sets of school effects for each of the grade-domains. They illustrate the strong affinity between the two sets of school effects.

Figure 4. School Effects: Betebenner SGP on the x-axis, OLS-based VAA on the y-axis



An important role of standardized tests is to help identify schools that perform exceptionally well, which can then be further examined to understand what drives their success, and schools that significantly under-perform, which can then be targeted for special assistance. Table 7 presents the frequencies of schools similarly classified by both methods within each cohort-domain as either significantly high or low performers, or as not significantly different from the median school, using a 95 percent confidence interval.[11] In all cohort-domains between 86 and 91 percent of schools are similarly classified by both methods; and none indicated as significantly above the median by one method is indicated as significantly below the median by the other. In general, more schools diverge significantly from the median in numeracy than in reading.

Table 7. Frequencies of schools similarly classified as significantly above or below the median

	Significantly Below	*Not significantly different*	*Significantly above*	*Sum*
Grade 3 to 5 Numeracy	9.4	64.9	12.0	86.3
Grade 3 to 5 Reading	2.3	79.8	4.5	86.6
Grade 7 to 9 Numeracy	11.2	66.9	11.2	89.3
Grade 7 to 9 Reading	4.6	83.5	3.5	91.6

To further examine the extent to which these three methods similarly identify exceptional schools we present in Table 8 the (un-named) schools that place among the ten top-ranked schools in grade 3 to 5 numeracy by either method, and the ten bottom-ranked schools, out of a total of 649 schools. They exhibit a great deal of overlap. Eight schools rank in the top ten by

[11] For the SGP effects we took as our confidence interval the values of the j^{th} and k^{th} order statistics where $j = ½ (n - 1.96\, n^{1/2})$ and $k = ½ (n + 1.96\, n^{1/2})$ (Conover, 1980).

both methods and seven schools rank in the bottom ten by both methods, and all schools in the top or bottom ten in one ranking are among the top or bottom thirty in the other ranking.

Table 8. Ten highest and lowest ranked schools by either method, numeracy, grade 3 to 5

	Regression-based	B-SGP based	Students in year level	Median prior score
Highest ranked (1 is highest rank)	1	1	28	431.2
	2	2	24	399.5
	3	3	27	384.7
	4	6	43	409.7
	5	28	28	455.4
	6	8	54	455.4
	7	10	116	389.6
	8	11	30	389.6
	9	7	27	389.6
	10	4	27	389.6
	24	5	37	399.5
	18	9	44	399.7
Lowest ranked (1 is lowest rank)	1	1	34	502.1
	2	4	30	442.9
	3	5	50	414.9
	4	2	35	442.9
	5	3	53	389.6
	6	14	23	455.4
	7	6	46	443.3
	8	22	30	431.2
	9	7	71	455.4
	10	13	65	389.6
	12	8	40	370.3
	25	9	27	399.5
	20	10	34	389.6

We conclude by examining the extent to which estimated school effects are sensitive to outlying student outcomes. We identify outlying student outcomes as student progress indicators that are in the top or bottom 2.5% of their respective distributions—regression residuals or growth percentiles. This is, of course, an arbitrary cutoff point. We have in mind outcomes that may be the result of ceiling or floor effects or of other sources of measurement error and hence do not reflect true progress in a way that is comparable to progress observed among other students. Extreme cases are students whose test score in the base year is higher than their score two years later, so they seem to have lost ground (as calibrated on an interval scale) and as a result their actual results are far below their expected outcomes. Conversely, there are students with extremely low scores in the base year who then appear to have made spectacular progress in the following two years. Either could reflect true progress but it is also possible that the very low score understates the student's achievement level in that year, perhaps because he or she did not feel well on the day of the test or possibly did not attach sufficient importance to doing well on the test. Of course, taking out the highest and lowest performers in a school when these truly reflect outstanding gains or losses will lead to *less* accurate estimates of school effects. Ideally one would want to examine the possibility of mis-measurement on a case-by-case basis, though this is clearly beyond the scope of the present paper.[12]

[12] As a step in this direction we recoded potential outliers in the data by combining low-frequency scores at the either end of the distribution, and in grades 7 and 9 numeracy, with the nearest high frequency scores. This affected less than 1% of observations. When we redid the calculations we found the effect of recoding on the aggregate indicators to be negligible.

To compare the sensitivity of regression-based and SGP-based school effects to the inclusion or exclusion of these outliers we recalculated the school effects omitting these students, and computed correlations between the school effects with and without these observations. These are presented in Table 10. Correlations are very high for both methods, with slightly higher correlations for the SGP-based rankings. This slight advantage may reflect the greater robustness of median-based methods to outliers or the limitations of linear models in dealing with floor and ceiling effects, where many of the outliers are found.

Table 9. Correlation of school effects with and without outliers

	Regression-based VAA	B-SGP
Grade 3-5, numeracy	0.96	0.98
Grade 3-5, reading	0.95	0.98
Grade 7-9, numeracy	0.95	0.97
Grade 7-9, reading	0.94	0.97

5. Concluding remarks

In this paper we compare two leading approaches to analyzing standardized test data linked over time: value-added analysis nased on least-squares regression (LS-VAA) and Betebenner's student growth percentile analysis (B-SGP). Both methods have a similar structure, first using prior scores to form individual normative expectations for current performance; then comparing

actual current scores to these normative expectations; and finally aggregating these comparisons to schools as an estimate of individual school effects on student progress.

We compare the two methods by applying them to NAPLAN test scores in two knowledge domains, numeracy and reading, for students in Victoria state schools progressing from grades 3 to 5 and 7 to 9 between 2008 and 2010, first comparing their characterization of individual progress and then comparing their estimates of school effects.

The conditional means produced by OLS regressions as a normative context for assessing current scores and the conditional medians produced by B-SGP for this purpose are nearly identical, with the exception of the upper and lower extremes of the distribution where we find a strong departure from linearity. The measures they produce assessing student progress, though defined differently—LS- VAA produces residuals where B-SGP produces conditional rankings—are also very closely aligned, with correlations between OLS residuals and B-SGP student percentiles between .94 and .96 for all cohort-domain pairs. In addition, the accuracy of predictions measured as the sum of squared differences between actual outcomes and conditional means for LS-VAA, and between actual outcomes and conditional medians for B-SGP, are nearly identical.

School effects derived from the two methods are also very similar, with correlations across schools within each cohort-domain ranging between 0.89 and 0.95. Identifying schools that are significantly better or worse than the median school under each method using a 95% confidence interval, and applying each method to classify schools as either significantly below the median, not significantly different, or significantly above the median, we find that between 86 percent and 91 percent of schools in each cohort-domain are similarly classified, and no school classified as significantly above the median by one method is identified as significantly below by the other.

The two methods are also very similar at the extremes of the distribution with a large degree of overlap between the lists of the ten highest-ranked schools of each method and between the lists of the ten lowest ranked schools. Finally we find that neither method is sensitive to outlying student observations: correlations of school effects with and without outliers range from .94 to .96 for LS-VAA school effects, and from .97 to .98 for B-SGP effects.

These statistical similarities suggest that despite the functional differences between least-squares the two methods—LS-VAA is geared primarily to supporting accountability of teachers and schools, B-SGP focuses on presenting a clear picture of student progress and a well-founded projection of future achievement—holding constant the underlying data, either method could be used for either purpose with very similar results.

References

Allen, Rebecca and Simon Burgess, 2011. Can School League Tables Help Parents Choose Schools? *Fiscal Studies* 32 (2): 245–261.

Ballou, Dale, 2009. Test Scaling and Value-Added Measurement. *Education Finance and Policy,* 4(4).

Ballou, Dale, 2008. Value-Added Analysis: Issues in the Economic Literature. Washington, DC: NRC: http://www7.nationalacademies.org/bota/VAM_Workshop_Agenda.html.

Ballou, Dale, William Sanders and Paul Wright, 2004. Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics* 29 (1): 37-65.

Betebenner, Damian W., 2009. Norm and Criterion-Referenced Student Growth. *Educational Measurement: Issues and Practice*, 28:42-51.

Betebenner, Damian W., 2011. *A technical overview of the student growth percentile methodology.* The National Center for the Improvement of Educational Assessment. http://www.nj.gov/education/njsmart/performance/SGP_Technical_Overview.pdf

Bond, Timothy N. and Kevin Lang, 2012. The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results. NBER WP 17960. Cambridge, MA: NBER.

Braun, Henry, 2005. Using student progress to evaluate teachers: A primer on value-added models. Princeton, NJ: Educational Testing Service Policy Information Center.

Briggs, Derek C. and Damian W. Betebenner, 2009. The invariance of measurement of growth and effectiveness to scale transformation. Paper presented at the 2009 NCME Annual Conference, San Diego, CA.

Briggs**,** D. C. , J.P. Weeks**,** and E. Wiley, 2008. The Sensitivity of Value-Added Modeling to the Creation of a Vertical Score Scale. *Education Finance and Policy, 4(4)*.

Castellano, Katherine Elizabeth and Andrew Ho, 2012. Contrasting OLS and quantile regression approaches to student "growth" percentiles. *Journal of Educational and Behavioral Statistics* Published online before print May 3, 2012 1076998611435413.

Conover, W.J., 1980. *Practical Nonparametric Statistics.* New York: John Wiley and Sons.

Dearden, L., J. Micklewright and A. Vignoles, 2011. The Effectiveness of English Secondary Schools for Pupils of Different Ability Levels. *Fiscal Studies* 32 (2): 225–244.

Ehlert, Mark, Cory Koedel, Eric Parsons and Michael Podgursky, 2012. Selecting growth measures for school and teacher evaluations. Harvard PEPG colloquium. Retrieved from: http://www.hks.harvard.edu/pepg/colloquia/2012-2013/signals_wp_ekpp_v7%20(2).pdf

Goldstein, H. and D.J. Spiegelhalter, 1996. League tables and their limitations: statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society: Series A* 159: 385–443.

Harris, Douglas, 2011. Value-added measures in education. Cambridge, MA: Harvard UP.

Helal, M., 2012. Measuring education effectiveness: Differential school effects and heterogeneity in value-added," Working paper, Melbourne Institute of Applied Economic and Social Research, University of Melbourne.

Mariano, Louis T., Daniel F. McCaffrey and J.R. Longwood, 2010. A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics*, 3(5):253-279.

McCaffrey, Daniel F., J.R. Lockwood, D. Koretz, T.A. Louis, and L. Hamilton, 2004. Models for Value-Added Modeling of Teacher Effects. *Journal of Educational and Behavioral Statistics* 29 (1): 67-101.

McCaffrey, Daniel F., Bing Han, and J.R. Lockwood, 2009.Turning student test scores into teacher compensations systems. In Springer, Mathew G. (ed.), *Performance Incentives: Their Growing Impact on American K-12 Education*. Washington, DC: Brookings, pp. 113-147.

National Research Council, 2010.Getting value out of value-added. Henry Braun, Naomi Chudowsky and Judith Koenig (eds.) Washington, DC: National Academies Press. Available at http://www.nap.edu/catalog/12820.html.

National Research Council, 2011. Incentives and test-based accountability in education. Michael Hout and Stuart Elliott (eds.). Washington, DC: National Academies Press. Available at: http://www.nap.edu/catalog.php?record_id=12521

OECD, 2008.Measuring improvements in learning outcomes: Best practices to assess the value-added of schools. Paris: OECD.

Ray, Andrew, Tanya McCormack, and Helen Evans, 2009.Value-added in English schools. *Education Finance and Policy, 4(4)*.

Reardon, S.F. and S.W. Raudenbush, 2009. Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4(4)*.

Sanders, W.L. and S.P. Horn, 2005. The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel Evaluation in Education*, 8(3).

Wright, Paul, 2010. An investigation of two nonparametric regression models for value-added assessment in education. SAS http://www.sas.com/resources/whitepaper/wp_16975.pdf.

Young, M.J., 2006. Vertical scales. In S.M. Downing and T.M. Haladyna (eds.), *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 469-485.