Estimating Teacher Effectiveness From Two-Year Changes in Students' Test Scores*

Andrew Leigh
Research School of Social Sciences
Australian National University
andrew.leigh@anu.edu.au
http://econrsss.anu.edu.au/~aleigh/

Abstract

Using a dataset covering over 10,000 Australian primary school teachers and over 90,000 pupils, I estimate how effective teachers are in raising students' test scores from one exam to the next. Since the exams are conducted only every two years, it is necessary to take account of the work of the teacher in the intervening year. Even after adjusting for measurement error, the resulting teacher fixed effects are widely dispersed across teachers, and there is a strong positive correlation between a teacher's gains in literacy and numeracy. Teacher fixed effects show a significant association with some, though not all, observable teacher characteristics. Experience has the strongest effect, with a large effect in the early years of a teacher's career. Female teachers do better at teaching literacy. Teachers with a masters degree or some other form of further qualification do not appear to achieve significantly larger test score gains. Overall, teacher characteristics found in the departmental payroll database can explain only a small fraction of the variance in teacher performance.

Keywords: education production function, teacher quality, panel data
JEL Codes: I21, J24, C23

* The paper has been improved as a result of comments from Steven Stillman, David Weimer, officers of the Department of Education, Training and the Arts, and seminar participants at the 2006 Labour Econometrics Workshop, the Australian National University, Queensland University of Technology, and the University of Melbourne. Elena Varganova provided outstanding research assistance. The support of the Australian Research Council's Discovery funding scheme (project number DP0665260) is acknowledged.

1. Introduction

In many occupations, it is relatively straightforward to estimate worker productivity. Standard proxies for output include billable hours for lawyers, value-added for builders, and research output for economists. But for school teachers, measuring output is considerably trickier. One commonly used measure of teacher effectiveness is expert assessment, in which an outside observer watches a teacher for some period of time, and forms a view as to his or her competence. However, since each observer only ever has the chance to see a relatively small number of teachers, the observer will typically find it difficult to compare the teacher with all other teachers, or to separate teacher-specific factors from other factors that may affect student achievement.

Given that children's test scores have been shown to be positively correlated with subsequent educational and labor market outcomes, exam results are often used as a measure of educational output.[1] Therefore, a natural measure of teacher productivity might be thought to be the average test scores of the children in that teacher's class. While this approach allows the use of a common benchmark for all teachers, it suffers from the problem that a large portion of the variance in children's test scores is determined by family background rather than by what is learned in schools (see eg. Coleman et al 1966).

This paper therefore seeks to estimate teacher output (or "teacher effectiveness") using changes in test scores from one test to the next.[2] Implementing such an approach requires panel data, in which teachers and students are observed over multiple years. Using a fixed effects regression, it is possible to separate student effects and teacher effects, and to thereby estimate something akin to the "value added" of a particular teacher.

[1] Test scores have been shown to be positively correlated with the high school graduation rate, future employment prospects and adult wages (Bishop 1991; Murnane, Willet and Levy 1995; Marks and Fleming 1998a, 1998b; Currie and Thomas 2001; Hanushek and Raymond 2002).

[2] In educational policy in Australia, the terms "teacher quality" or "teacher effectiveness" is in common parlance, but is typically associated with pedagogy, work practices, and professional development programs: focused on improving the quality of teaching, given the current teaching workforce (see eg. Australian College of Educators 2001, and the reports of the Teacher Quality and Educational Leadership Taskforce). By contrast, my measure of teacher effectiveness is individual-specific, and does not change over the brief period in which the students were observed.

By contrast with approaches that investigate the correlation between student and teacher characteristics in a single cross-section, the use of panel data makes it possible to take account of the fact that teachers are not randomly assigned to students. This is true both across schools (teachers may choose to work at a particular school because of the makeup of the student body), and within schools (principals may assign the most effective teachers to the most gifted or struggling students). Panel data take account of this issue by including out a student fixed effect, thereby making it possible to compare the performance of the same student under different teachers.

A similar strategy to that implemented in this paper has been carried out in three recent US studies. Using data from Texas, Rivkin, Hanushek and Kain (2005) estimate a fixed effects model on a population of over half a million students. Their dataset allows them to identify the school and grade for each teacher and student. For schools with only one teacher per grade, this allows them to match teachers and students perfectly, while for other schools, they are able to match groups of teachers with groups of students. Rivkin, Hanushek and Kain find that differences between teachers explain about 15 percent of the measured variance in student test scores. In both reading and mathematics, a one standard deviation increase in teacher effectiveness leads to an increase in student achievement of around one-tenth of a standard deviation. The authors also explore the issues of qualifications and turnover, concluding that teacher qualifications explain little of the variance in teacher effectiveness, and that those teachers who leave the profession are not substantially different from those who remain.

Similar research by Rockoff (2004) uses data from two school districts in New Jersey. While Rockoff's sample comprises only about 10,000 students, he has the significant advantage that he is able to precisely match students to teachers. Rockoff finds significant variation in teacher effectiveness, with a point estimate similar to Rivkin, Hanushek and Kain: moving one standard deviation up the distribution of teacher fixed effects raises students' reading and math test scores by about one-tenth of a standard deviation on the national scale. A similar study by Aaronson, Barrow and Sander (2007), using data from Chicago high schools, find that a one standard

deviation increase in math teacher effectiveness over a full year raises student test scores by 0.15 standard deviations.

Outside the United States, relatively little research has been carried out on the measurement of teacher effectiveness.[3] One of the main challenges is that standardized tests are often not administered annually. For example, elementary school pupils are typically tested in grades 3, 5 and 7 in Australia, ages 5, 7 and 11 in England, ages 8 and 11 in France, and grades 2, 5 and 7 in Sweden (O'Donnell 2004). The estimation of teacher fixed effects models with biennial data is therefore an issue of considerable policy relevance.

This paper uses data from the state of Queensland, Australia, where standardized tests are conducted every two years. With over 90,000 primary school pupils in grades 3 to 7 between 2001 and 2004, it is possible to estimate the teacher fixed effects for over 10,000 teachers. To preview the results, I find that the teacher fixed effects are jointly significant, and highly dispersed. Moving from a teacher at the 25th percentile to a teacher at the 75th percentile would raise test scores by one-seventh of a standard deviation. I find that teacher experience is positively correlated with teacher effectiveness, but find no positive effects of teacher qualifications on test scores. Female teachers do better at teaching literacy. Overall, however, these factors account for less than one-hundredth of the variation between teachers. Most of the differences between teachers are due to factors not captured in the payroll database.

The remainder of this paper is organized as follows. Section 2 outlines the methodology, and estimates a teacher fixed effects model. Section 3 analyses the teacher fixed effect terms, to see how much of the variation between teachers can be

[3] In Australia, the closest study to this one is Hill and Rowe (1996), who use data from 13,700 Victorian primary and secondary school children to estimate the fraction of test score variance within classes, and within schools. They conclude that variance at the class/teacher level constitutes 37-54 percent of measured variance, while school-level variance constitutes just 4-8 percent of total variance. A similar study focusing on year 12 students found that class/teacher effects consistently accounted for 59 percent of the residual variance in student achievement, compared with 5 percent at the school level (Rowe 2000; Rowe, Turner and Lane 1999, 2002). Yet a significant drawback of these studies (unavoidable given the data available to the researchers) is that they are unable to take account of the non-random allocation of students across schools, and teachers across classrooms. As a result, one cannot know whether classroom-level variance is high because there are substantial differences in teacher quality, or because schools tend to sort students into classes based on ability.

explained by qualifications and demographic characteristics. The final section concludes.

2. Estimating Teacher Fixed Effects With Biennial Tests

This study uses de-identified microdata for primary students between grades 3 and 7 who attended government schools in the state of Queensland during the years 2001-2004. Queensland administers standardized literacy and numeracy tests to all pupils in grades 3, 5 and 7. Since the focus is on differences from one test to the next, I restrict the sample to students who completed two tests. Due to data problems with one cohort, the final sample consists of three cohorts of students, depicted in Table 1.[4]

Table 1: Cohorts Used in the Study
Test years marked in italics.

Year	**Cohort 1**	**Cohort 2**	**Cohort 3**
2001	*Grade 3*		*Grade 5*
2002	Grade 4	*Grade 3*	Grade 6
2003	*Grade 5*	Grade 4	*Grade 7*
2004		*Grade 5*	
Sample Size			
Literacy test	30,604	31,208	30,658
Numeracy test	30,715	31,278	30,826

Note: Sample size is in the first year only. Not all students are observed in all years.

In order to estimate the relationship between teacher characteristics and changes in student test scores, it is necessary to match data from four different files.

(i) Using a dataset of test scores, I use education department identifier codes and students' birth dates to match students' performance in one test with their performance in the test taken two years later.

(ii) Using a dataset of student assignments to roll classes, I use education department identifier codes and students' birth dates to match students to a particular classroom in each of the three years that they appear in the sample.

(iii) Using a dataset of teacher assignments to roll classes, I use roll class identifiers and school codes to match teachers to classrooms.

(iv) Using a dataset of teacher payroll information, I use teacher payroll identifiers to match teachers to their age, experience, qualifications and gender.

[4] The test scores provided by DETA for students who took the grade 7 test in 2004 were missing education department identifier codes.

Because some students move between grades, are absent on the day of the test, or have their birthdates mis-coded in the dataset, I am only able to make an exact match for about three-quarters of students in the sample. From an initial cohort of around 40,000, the sample sizes in Table 1 are around 30,000-32,000.

The timing of tests in Australia also introduces complications. Previous papers that estimate teacher fixed effect models (such as Rivkin, Hanushek and Kain 2005 and Rockoff 2004) use data from elementary school exams that are administered annually, at the end of the school year. As a result, any change from one test to the next can be attributed to only one teacher (assuming no teacher turnover during the year).

By contrast, Queensland (like other Australian states and territories) administers its statewide standardized test biennially. Thus the question arises of how teachers in the intervening year should be treated. The two most plausible approaches are: (a) ignore the intervening year altogether, or (b) create an assumed test score in the intervening year, which lies at the midpoint of the other two tests. In this section, I present both methods, the results of which turn out to be quite similar. To maximize sample size, I therefore use the interpolation method in the following section.

A second complication is that tests are administered just after the middle of the school year (the school year runs from January to December, and the tests are administered in August). In the case of a child who takes tests in the middle of grade 3 and the middle of grade 5, it is therefore possible that the grade 3 teacher contributes to both tests. Under most plausible assumptions, this will introduce only attenuation bias into estimates of the teacher fixed effects terms. To the extent that teachers focus their attention on the test administered in their year, or the test is based on material taught in that grade and the preceding grade, the attenuation bias introduced by using mid-year tests will be smaller than otherwise.

I use the results of 12 tests – literacy and numeracy exams administered to three cohorts of students at two grade levels. Although the tests are scaled so as to be comparable over time and across grades, I standardize each of the tests to a mean of

zero and a standard deviation of unity.[5] Thus the average student has a test score of zero, and the average change in the relative distribution of student test scores is zero. Naturally, this does not mean that the average student learns nothing between tests, but that the average student's *relative position* in the distribution remains unchanged between tests. A student who is 0.5 standard deviations above the mean is performing at about the same level as the typical child in the next grade.[6]

I then estimate the following regression:

$$X_{ijsgt} = \alpha + \beta_j T_j + C_{jsgt} + \Psi_{gt} + \Omega_s + \Pi_i + \varepsilon_{ijsgt} \quad (1)$$

X_{ijsgt} is the literacy or numeracy test score of individual i, taught by teacher j, in school s, grade g, and calendar year t. The main focus is on β_j, the coefficients on the teacher fixed effect terms T_j. Other controls are class size C_{jsgt}, grade*calendar year fixed effects Ψ_{gt}, school fixed effects Ω_s, and student fixed effects Π_i. ε_{ijsgt} is a normally distributed error term.

An important advantage of this methodology is that focuses not on students' performance in a single test, but on the difference between their performance in one test and another. This helps to deal with one of the most common criticisms of exams as a measure of school performance: that differences between students are determined primarily by children's home environment, rather than what they learn in the classroom.[7]

[5] Such a rescaling has two advantages. First, it makes the coefficients more readily interpretable. Second, it avoids the problem that the dispersion of test scores tends to change systematically across grades (falling for literacy, and rising for numeracy). Re-estimating the results using the raw scores makes no substantive difference to the results.

[6] This calculation uses the fact that the original scores are designed to be comparable across grades and years. In literacy, a student must score 0.57 standard deviations above the mean to be equivalent to a child in the next grade. In numeracy, a student must score 0.48 standard deviations above the mean to be equivalent to a child in the next grade.

[7] It is possible that a student's home background affects not only the level of her scores, but also her gain from one test to the next. Whether students at the bottom of the distribution tend to have larger or smaller gains than those at the top of the distribution will depend primarily on the way in which the test is scaled. Ideally, one might wish to include two student fixed effects – one for the level, and another for the gain. However, the data provided to me by DETA contains only two observations per student, which makes it possible to include only a level fixed effect for each student.

Setting the standard deviation of the student test score distribution to one gives the teacher fixed effects a straightforward interpretation. For example, a teacher with a fixed effect of one raises her students' test scores on average by one standard deviation, relative to all other teachers. Naturally, because the average change in student test scores is zero, the average teacher fixed effect is also zero (ie. students of the average teacher maintain their position in the relative student test score distribution).

The results of the student-level regression are shown in Table 2.[8] These are not in themselves particularly informative; what matters most is that the teacher fixed effects are jointly significant. Omitting non-test years (Panel A), the value of the F-test of the hypothesis that the teacher fixed effects terms are equal to zero is 2.84 for literacy and 3.98 for numeracy. Linearly interpolating test scores in non-test years (Panel B), the value of the F-test of the hypothesis that the teacher fixed effects terms are equal to zero is 2.93 for literacy and 4.10 for numeracy. In all cases, these F-tests easily reject the null that there are no systematic differences between teachers.

The class size coefficients are positive (and statistically significant in the case of literacy tests). On its face, this suggests that larger classes produce better literacy outcomes, but given the presence of nonrandom sorting of students across differently sized classes, readers are urged not to draw any causal inference from this coefficient.

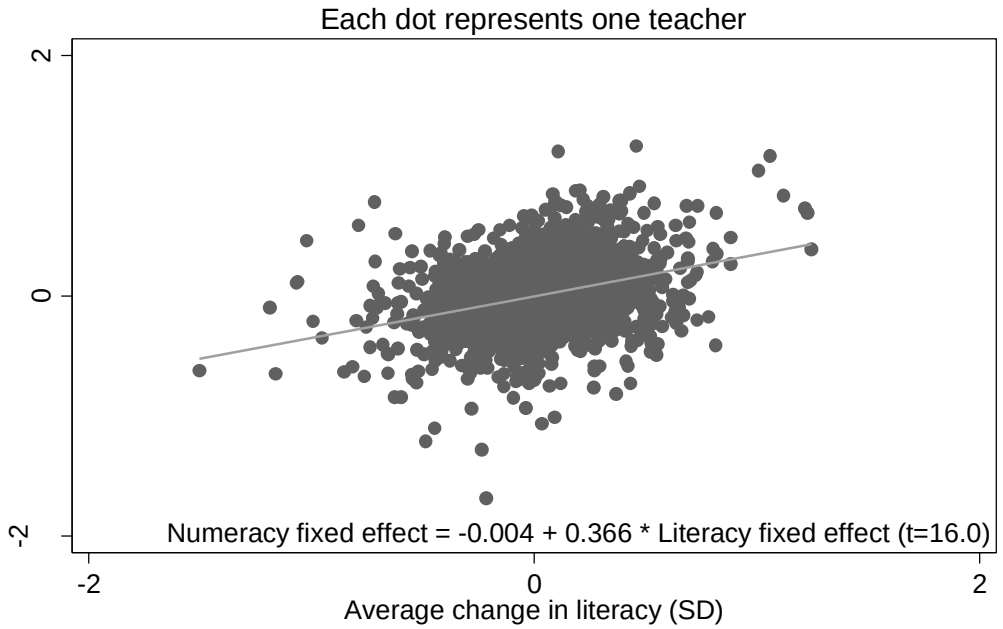
[8] Computationally, the student fixed effects and school fixed effects are estimated by de-meaning the data, since at the time of writing, I was unable to obtain sufficient computing power to run a regression with this many fixed effects. For a detailed discussion of the various approaches used to estimate fixed effects models in the presence of computational constraints, see Abowd, Kramarz and Margolis (1999).

Table 2: Estimating teacher fixed effects from panel data		
	(1)	(2)
Dependent variable:	**Literacy test**	**Numeracy test**
Panel A: Dropping Non-Test Years		
Class size	0.0009***	-0.0003
	[0.0003]	[0.0003]
Grade*Calendar Year fixed effects	Yes	Yes
Individual fixed effects	Yes	Yes
School fixed effects	Yes	Yes
Teacher fixed effects	Yes	Yes
Observations (students*years)	185,188	186,014
Number of students	95,914	96,001
Number of teachers	9238	9248
Number of schools	1058	1058
F-test for joint significance of	2.84	3.98
teacher fixed effect terms	[P=0.0000]	[P=0.0000]
Panel B: Interpolating Non-Test Years		
Class size	0.0005***	-0.0002
	[0.0002]	[0.0002]
Grade*Calendar Year fixed effects	Yes	Yes
Individual fixed effects	Yes	Yes
School fixed effects	Yes	Yes
Teacher fixed effects	Yes	Yes
Observations (students*years)	273,236	274,783
Number of students	95,921	96,008
Number of teachers	10,749	10,821
Number of schools	1058	1058
F-test for joint significance of	2.93	4.10
teacher fixed effect terms	[P=0.0000]	[P=0.0000]

Notes: ***, ** and * denote statistical significance at the 1%, 5% and 10% levels respectively. Standard errors in brackets. In Panel B, grade 4 students are assigned the average of their grade 3 and 5 tests; and grade 6 students are assigned the average of their grade 5 and 7 tests.

Not surprisingly, the teacher fixed effects for literacy and numeracy are highly correlated. Using the results from Panel B, Figure 1 shows a plot of the two fixed effects for each teacher in the sample. For the most part, teachers whose pupils have above-average numeracy gains also have above-average literacy gains; while teachers whose pupils have below-average numeracy gains also have below-average literacy gains.

Figure 1: Are Good Numeracy Teachers Also Good Literacy Teachers?



The dispersion of the teacher fixed effects terms provides a measure of the dispersion of teacher performance across Queensland primary schools. However, because the teacher fixed effects are measured with error, the true variance of teacher performance will be narrower than the distribution of the teacher fixed effects terms. Using the maximum likelihood shrinkage estimator described in Rockoff (2004), I use the teacher fixed effects terms and their associated standard errors to estimate the true standard deviation of teacher effectiveness. These estimates are set out in Table 3. Using the shrinkage estimator, the standard deviation on the teacher fixed effects terms falls to around 0.12-0.13 when non-test years are dropped, and to around 0.09-0.11 when non-test years are interpolated. This indicates a very similar level of dispersion across teachers in Queensland primary schools as has been observed across schools in New Jersey and Texas.

Table 3: Standard Deviation of Teacher Fixed Effect Terms

	Raw	Adjusted
Panel A: Dropping Non-Test Years		
Literacy	0.199	0.119
Numeracy	0.191	0.137
Panel B: Interpolating Non-Test Years		
Literacy	0.153	0.092
Numeracy	0.158	0.107

Note that even adjusting the dispersion, there is a wide distribution of teacher fixed effects. Taking the standard deviation of teacher fixed effects to be 0.1, the results above therefore suggest that moving from a teacher at the 25th percentile to a teacher at the 75th percentile would raise test scores by one-seventh of a standard deviation. Recalling that a 0.5 standard deviation increase in test scores is equivalent to a full year's learning, this implies that a 75th percentile teacher can achieve in three-quarters of a year what a 25th percentile teacher can achieve in a full year.

Moving from a teacher at the 10th percentile to a teacher at the 90th percentile would have even more dramatic effects, raising test scores by one quarter of a standard deviation. This implies that a teacher at the 90th percentile can achieve in half a year what a teacher at the 10th percentile can achieve in a full year. Moreover, since the teacher fixed effects are estimated from a regression that includes school fixed effects, it is possible that the true dispersion of teacher effectiveness is even wider than these results suggest.

To make the above simulations more concrete, note that the test score gap between Indigenous and Non-Indigenous students in Australia is approximately one standard deviation (see eg. Rothman 2002). Assuming the gap is just as large in Queensland primary schools, this implies that Indigenous students perform approximately two grades below their non-Indigenous counterparts.[9] Assuming that the impact of having a more effective teacher persists over time, and that Indigenous children typically get teachers at the 25th percentile, these results imply suggests the black-white test score gap in Australia could be closed in seven years by giving all Indigenous pupils teachers at the 75th percentile.

3. How Much Can Variation in Teacher Effectiveness be Explained by Demographics?

Having derived a teacher fixed effect for each teacher in the sample, it is possible to ask the question: how much of the variation between teachers can be explained by characteristics such as gender, age, experience, and qualifications? This question has

[9] The dataset that I have been supplied by DETA does not include any student demographic characteristics.

important policy ramifications, since the uniform salary schedules that operate in Australian public schools are based exclusively on experience and qualifications. To the extent that these factors are good proxies for productivity, such a system will appropriately remunerate the teaching workforce. However, if experience and qualifications do not explain a large portion of the variation between teachers, this suggests that the uniform salary schedules may be overly rigid.

Table 4 sets out the characteristics of the 10,662 teachers in the sample (for this part of the paper, I drop teachers with missing demographics or fixed effects). Around 10 percent have a masters degree or some further qualification. (Since the 1980s, registered teachers in Queensland public schools have been required to complete at least 4 years of tertiary training. This category covers those who have done more than the minimum requirement to be registered, such as an honors degree, a masters, a doctorate, or a second degree.) The share of teachers who are female is 77 percent, the average age is 40, and the average number of years of experience is 13. The Department of Education, Training and the Arts (DETA) also provides a "suitability rating" for 6379 teachers, or about two-thirds of the sample. Teachers receive a rating of 1 ("outstanding applicants"), 2 ("quality applicants"), 3 ("satisfactory applicants"), or 4 ("eligible for temporary/casual employment").[10] Across the sample, 73 percent of teachers were rated in the top category, 18 percent in the second-highest category, 10 percent in the third-highest category. Only one teacher in the sample received a rating in the lowest category, so I combine categories 3 and 4.

Table 4: Unweighted Summary Statistics for Teacher Characteristics

Variable	Obs	Mean	Std. Dev.
Masters Degree or Other Further Qualification	10662	0.101	0.301
Female	10662	0.772	0.419
Age	10662	40.067	10.452
Experience	10662	13.412	11.015
DETA Rating=1	6379	0.726	0.446
DETA Rating=2	6379	0.177	0.382
DETA Rating=3 or 4	6379	0.097	0.296

In Table 5, I show the results of regressing the teacher fixed effect terms on the various observable characteristics that are available from the DETA payroll database.

[10] Teachers who have not yet been rated are given a suitability rating of "T4". Since this does not reflect the department's assessment of their competence, I code it as missing.

Panel A shows the results using the teacher's fixed effect based on changes in literacy scores, while Panel B uses teacher fixed effects based on changes in numeracy scores.

Before discussing the particular coefficients, it is worth noting that while several teacher characteristics are systematically related to teacher fixed effects, very little of the variance between teachers can be explained by the factors in the DETA payroll database. As the results in Tables 3 and 4 show, there are large gaps between teachers. However, as the R-squared statistics in Tables 5 and 6 indicate, the combination of qualifications, gender, age, experience and the DETA ratings explain less than 1 percent of the variation between teachers.

For both literacy and numeracy, I find that teachers with a masters degree or some other further qualification obtain lower test score gains than teachers without these additional qualifications. This effect is statistically significant with or without additional demographic controls. The absence of a positive effect of teacher qualifications on teacher performance is consistent with US studies (Rivkin, Hanushek and Kain 2005 and Rockoff 2004), which also find no impact of having a masters degree. However, it should be noted that my estimates – and those from the US – are based upon comparing those teachers who chose to obtain masters degrees with those who did not. It is entirely plausible that masters degrees have a positive impact on student test score gains, but that there is some negative selection into masters programs. A preferable estimation strategy would be to observe teachers before and after they obtain a masters degree; but this is not feasible with the present dataset.

There appears to be some significant effect of teacher gender on student test score gain. In particular, female teachers have larger test score gains in literacy, a result that is robust to controlling for age, experience and qualifications. In numeracy, the female coefficient is negative, but insignificant and small in magnitude.

Age and experience are positively related to student test score gains. In the literacy specification, including both age and experience causes the coefficients to become statistically insignificant. In the numeracy specification, the effects of age and

experience are larger in magnitude than for literacy, and the effects remain statistically significant when both are included in the regression.

Note that with only a short panel, I am unable to separate cohort effects from age effects. In the present situation, this may be important, given that the academic aptitude of new teachers in Australia was significantly lower in the early-2000s than in the early-1980s (Leigh and Ryan 2006). Assuming a teacher's academic aptitude is positively correlated with his or her teacher fixed effect, this secular decline in teacher aptitude will cause an upward bias in the age coefficient.

Table 5: Student Test Score Gain and Teacher Characteristics
Dependent Variable is the teacher fixed effect

	(1)	(2)	(3)	(4)	(5)
			Panel A: Literacy		
Masters	-0.0084**				-0.0068*
	[0.0039]				[0.0039]
Female		0.0133***			0.0142***
		[0.0025]			[0.0026]
Age			0.0003***		0.0002
			[0.0001]		[0.0002]
Experience				0.0002**	0.0002
				[0.0001]	[0.0001]
R-squared	0.0005	0.0028	0.0008	0.0007	0.0044
Teachers	10662	10662	10662	10662	10662
			Panel B: Numeracy		
Masters	-0.0109***				-0.0078**
	[0.0038]				[0.0039]
Female		-0.0034			-0.0014
		[0.0027]			[0.0028]
Age			0.0007***		0.0003**
			[0.0001]		[0.0002]
Experience				0.0006***	0.0004**
				[0.0001]	[0.0002]
R-squared	0.0008	0.0002	0.0035	0.0039	0.0047
Teachers	10662	10662	10662	10662	10662

Notes: ***, ** and * denote statistical significance at the 1%, 5% and 10% levels respectively. Robust standard errors in brackets. Each observation is a teacher fixed effect (derived from the specifications set out in Table 2). Estimates are weighted by the number of students taught by each teacher.

In Table 6, I estimate the effect of the DETA rating, which is available for about two-thirds of the teachers in the sample. By comparison with teachers rated 3 or 4 (the two lowest ratings), teacher rated 1 or 2 produce higher test score gains. However, the

positive relationship between the DETA rating and value-added is only statistically significant for literacy, where DETA ratings have predictive power even controlling for gender, qualifications, age and experience. Overall, it seems that better-rated teachers do indeed achieve higher student test score gains on the literacy test.

Table 6: Student Test Score Gain and Education Department Ratings
Dependent Variable is the teacher fixed effect

	(1)	(2)
	Panel A: Literacy	
Rating=1	0.0091*	0.0138***
	[0.0049]	[0.0050]
Rating=2	0.0073	0.0110*
	[0.0057]	[0.0058]
Masters		-0.0061
		[0.0045]
Female		0.0082**
		[0.0037]
Age		0.0001
		[0.0002]
Experience		0.0005**
		[0.0002]
R-squared	0.0006	0.0037
Teachers	6379	6379
	Panel B: Numeracy	
Rating=1	0.0009	0.0086
	[0.0053]	[0.0057]
Rating=2	0.0037	0.0082
	[0.0062]	[0.0063]
Masters		-0.0034
		[0.0046]
Female		-0.0066*
		[0.0038]
Age		0.0003*
		[0.0002]
Experience		0.0007**
		[0.0003]
R-squared	0.0001	0.0047
Teachers	6379	6379

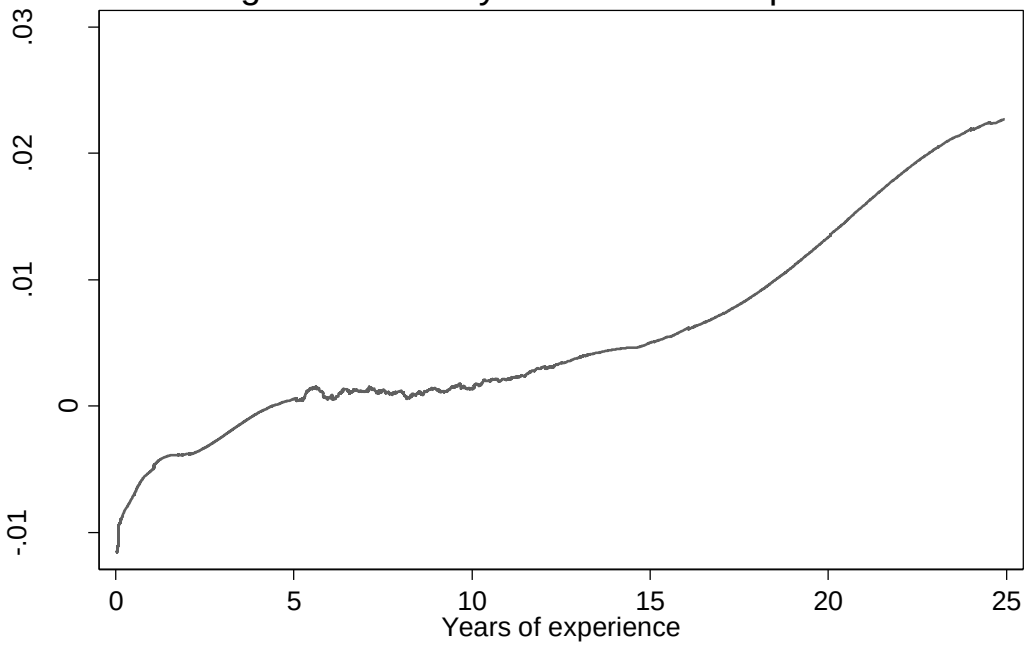
Notes: ***, ** and * denote statistical significance at the 1%, 5% and 10% levels respectively. Robust standard errors in brackets. Each observation is a teacher fixed effect (derived from the specifications set out in Table 2). Estimates are weighted by the number of students taught by each teacher. The DETA rating ranges from 1 to 4, with 3-4 being the excluded category from the regressions (only one teacher in the sample is rated 4).

Since Tables 5 and 6 only include experience as a linear term, Figures 2 and 3 test whether there is a nonlinear relationship between experience and student test score

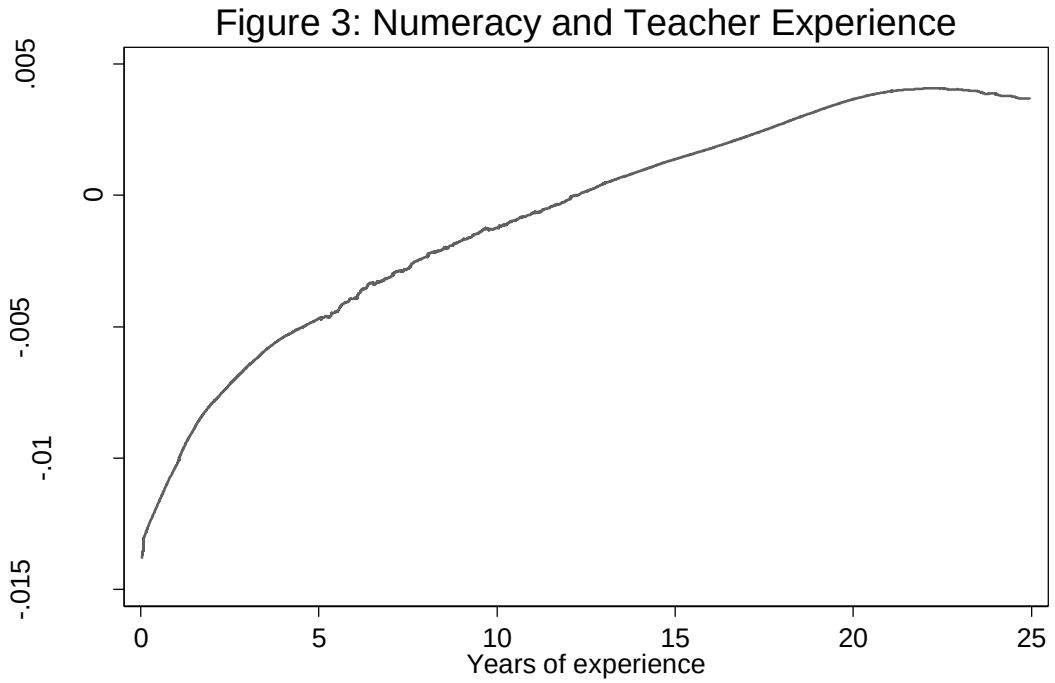
gain. Both charts are based upon locally weighted regression of teacher fixed effects on experience.

For both literacy and numeracy, there appears to be a significant effect of experience in the early years. Compared to teachers with 10 years of experience, novice teachers have test score gains that are 1/100ths of a standard deviation lower in literacy, and nearly 2/100ths of a standard deviation lower in numeracy. While the experience effect increases for those with more than 10 years of experience, the marginal effect of another year of experience declines. The effects of experience are larger for literacy than for numeracy.

Figure 2: Literacy and Teacher Experience



Note: Results are based on locally weighted regressions.



Note: Results are based on locally weighted regressions.

4. Conclusion

This paper has shown how to estimate a measure of teacher performance by using panel data with two test scores per student; parsing out the effects of family background by including student fixed effects. Rather than looking at which teachers have students that are at the top or bottom of the distribution, this approach effectively asks which teachers have students who moved up or down the distribution from one test to the next.

So far as I am aware, this is the first paper outside the United States to implement this empirical strategy, and the first to estimate a teacher fixed effects model using biennial data. While US tests are conducted annually (making them readily usable for estimating teacher fixed effects models), Australian tests are conducted only every second school year. However, this paper demonstrates that this is not an insurmountable obstacle, and that by either dropping teachers in the middle year, or interpolating test scores in intervening years, it is possible to observe the effects of teachers on student test score gains.

The differences between the best and worst teachers in Queensland are considerable. After adjusting for measurement error in estimating the teacher fixed effects terms, I find that the standard deviation of teacher fixed effects is around 0.1, similar to estimates from other studies in the United States. This suggests that moving from a teacher at the 25th percentile to a teacher at the 75th percentile would raise test scores by one-seventh of a standard deviation. In terms of literacy and numeracy test scores, a 75th percentile teacher can achieve in three-quarters of a year what a 25th percentile teacher can achieve in a full year; while a 90th percentile teacher can achieve in half a year what a 10th percentile teacher can achieve in a full year.

These results suggest that raising teacher effectiveness may be at least as cost-effective policy reform as reducing class sizes. An oft-cited upper bound of the effects of class size reductions on test scores is Krueger (1999), whose estimates suggest that reducing class sizes by one-sixth would boost test scores by 0.11 standard deviations. It is not unreasonable to think that an equivalent expenditure – a one-sixth increase in teacher salaries – might lead to a one standard deviation increase in teacher effectiveness (raising the average teacher to what is now the 84th percentile), thus producing an equivalently large increase in student achievement.

There are also two other reasons why focusing on the quality of the teaching workforce appears more attractive than class size reductions. First, it is possible that the benefits of class size reductions are considerably smaller than the estimates of Krueger (1999). For example, Hoxby (2000) and Hanushek (1998) find zero or negligible benefits of across-the-board class size reductions. Second, it is quite plausible that large-scale class-size reductions have the effect of lowering teacher quality, particularly in disadvantaged schools (Jepsen and Rivkin 2002).

The results from this paper also shed light on the extent to which uniform pay schedules, which reward teachers based solely upon qualifications and experience, capture productivity differences between teachers. It is certainly true that some of the variation between teachers can be explained by demographic factors. In both literacy and numeracy, more experienced teachers have higher test score gains. The experience effect is large in the early years of a teacher's career, and thereafter appears to be larger for numeracy than literacy (though it is possible that this also

reflects a decline in the aptitude of new teachers over time). I find suggestive evidence that students with female teachers do better in literacy, but no evidence that students do better if their teachers have higher formal qualifications. And the DETA rating does seem to capture some differences between teachers, even holding constant other characteristics.

Yet while there are some systematic patterns, 99 percent of the variation in teacher performance remains unexplained by differences in teacher demographics. This suggests that uniform pay schedules are indeed only picking up a small portion of the differences in test score gains across teachers. Assuming test score gains are an important measure of educational output, these results suggest that it may be worth considering alternative salary structures, as a means of attracting and retaining the best teachers.

References

Aaronson, D., Barrow, L., and Sander, W. (2007) "Teachers and Student Achievement in the Chicago Public High Schools", *Journal of Labor Economics*, 25(1): 95-135.

Abowd, J.M, Kramarz, F. and Margolis, D.N. (1999). "High Wage Workers and High Wage Firms," *Econometrica*, 67(2): 251-333.

Australian College of Educators (2001) "Teacher Standards, Quality and Professionalism: Towards a Nationally Agreed Framework", ACE Statement

Bishop, J. (1991) "Achievement, Test Scores and Relative Wages" in M.H. Kosters (ed) *Workers and Their Wages*, Washington DC, AEI Press, 146-186

Coleman, J.S., Campbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. (1966) *Equality of Educational Opportunity*. Washington, D.C.: U.S. Government Printing Office

CREDO (2002) "The Future of California's Academic Performance Index" Hoover Institution, Stanford University, mimeo

Currie, J. and Thomas, D. (2001) "Early Test Scores, Socioeconomic Status, School Quality and Future Outcomes" *Research in Labor Economics*, 20: 103-132

Hanushek, E. A. (1998) "The Evidence on Class Size", W. Allen Wallis Institute of Political Economy Occasional Paper Number 98-1, University of Rochester, Rochester, NY

Hanushek, E.A. and Raymond, M.E. (2002) "Improving Educational Quality: How Best to Evaluate Our Schools?", Paper prepared for Education in the 21st Century: Meeting the Challenges of a Changing World, Federal Reserve Bank of Boston, June 19-21

Hill, P.W., and Rowe, K.J. (1996). Multilevel modeling in school effectiveness research. *School Effectiveness and School Improvement*, 7 (1): 1-34.

Hoxby, C.M. (2000) "The effects of class size on student achievement: New evidence from population variation," *Quarterly Journal of Economics*, 115(4): 1239-1285.

Jepsen, C. and Rivkin, S.G., (2002) "What is the Tradeoff Between Smaller Classes and Teacher Quality?" NBER Working Paper No. 9205. NBER: Cambridge, MA.

Krueger, A. (1999) "Experimental Evidence of Education Production Functions," *Quarterly Journal of Economics*, 114(2): 497-532

Leigh, A. and Ryan, C. (2006) "How and Why has Teacher Quality Changed in Australia?" Australian National University CEPR Discussion Paper 534, Canberra, ACT: ANU.

Marks, G. and Fleming, N. (1998a) *Factors Influencing Youth Unemployment in Australia 1980-1994*, Longitudinal Surveys of Australian Youth Research Report No. 7, Melbourne: ACER

Marks, G. and Fleming, N. (1998b) *Youth Earnings in Australia 1980-1994: A Comparison of Three Youth Cohorts*, Longitudinal Surveys of Australian Youth Research Report No. 8, Melbourne: ACER

Murnane, R.J., Willet, J.B., and Levy, F. (1995) "The Growing Importance of Cognitive Skills in Wage Determination" *Review of Economics and Statistics* 77: 251-266

O'Donnell, S. 2004. *International Review of Curriculum and Assessment Frameworks Comparative tables and factual summaries*, 14th edition, Qualifications and Curriculum Authority and National Foundation for Educational Research, London, UK.

Rivkin, S.G., Hanushek, E.A., and Kain, J.F. (2005) "Teachers, Schools, and Academic Achievement", *Econometrica* 73(2): 417-458.

Rockoff, J.E. (2004) "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data", *American Economic Review*, 94(2): 247-252.

Rothman, S. (2002). "Achievement in Literacy and Numeracy by Australian 14-Year-Olds, 1975-1998" Research Report Number 29, ACER: Melbourne.

Rowe, K. (2000). *Celebrating coeducation? Maybe, but not necessarily for academic achievement*! *An examination of the emergent research evidence*. Invited keynote address presented at the Second National Conference on Co-education, Kinross Wolaroi School, Orange, New South Wales, April 16-19, 2000.

Rowe, K.J., Turner, R., and Lane, K. (1999). *The 'myth' of school effectiveness: Locating and estimating the magnitudes of major sources of variation in students' Year 12 achievements within and between schools over five years.* Paper presented at the 1999 AARE-NZARE Joint Conference of the Australian and New Zealand Associations for Research in Education, Melbourne Convention Centre, November 29 – December 2, 1999

Rowe, K.J., Turner, R., and Lane, K. (2002). Performance feedback to schools of students' Year 12 assessments: The *VCE Data Project*. In A.J. Visscher and R. Coe (Eds.), *School improvement through performance feedback* (pp. 163-190). Lisse, The Netherlands: Swetz & Zeitlinger.