Economics of Education Review 29 (2010) 480–488

Contents lists available at ScienceDirect

Economics of Education Review

journal homepage: www.elsevier.com/locate/econedurev

Estimating teacher effectiveness from two-year changes in students' test scores

Andrew Leigh*

Research School of Economics, Australian National University, ACT 0200, Australia

ARTICLE INFO

Article history:
Received 15 April 2008
Accepted 4 August 2009

JEL classification:
I21
J24
C23

Keywords:
Educational economics
Educational finance
Efficiency
Productivity

ABSTRACT

Using a dataset covering over 10,000 Australian school teachers and over 90,000 pupils, I estimate how effective teachers are in raising students' test scores. Since the exams are biennial, it is necessary to take account of the teacher's work in the intervening year. Even adjusting for measurement error, the teacher fixed effects are widely dispersed, 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 impact, particularly in the early years of a teacher's career. Female teachers do better at teaching literacy. Teachers with a master's 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 explain only a small fraction of the variance in teacher performance.

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, e.g. 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. 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

* Tel.: +612 6125 1374; fax: +612 6125 0182.
E-mail address: andrew.leigh@anu.edu.au.

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

0272-7757/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.econedurev.2009.10.010

effects, and to thereby estimate something akin to the "value-added" of a particular teacher.

By contrast to 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 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) estimated a fixed effects model on a population of over half a million students. Their dataset allowed them to identify the school and grade for each teacher and student. For schools with only one teacher per grade, this allowed them to match teachers and students perfectly, while for other schools, they were able to match groups of teachers with groups of students. Rivkin et al. found that differences between teachers explained about 15% of the measured variance in student test scores. In both reading and mathematics, a one standard deviation increase in teacher effectiveness led to an increase in student achievement of around one-tenth of a standard deviation. The authors also explored the impact of qualifications and turnover, concluding that teacher qualifications explained little of the variance in teacher effectiveness, and that those teachers who left the profession were not substantially different from those who remained.

Similar research by Rockoff (2004) used data from two school districts in New Jersey. While Rockoff's sample comprised only about 10,000 students, his study had the advantage that he was able to precisely match students to teachers. Rockoff found significant variation in teacher effectiveness, with a point estimate similar to Rivkin et al.: moving one standard deviation up the distribution of teacher fixed effects raised students' reading and mathematics test scores by about one-tenth of a standard deviation on the national scale. At the high school level, a study by Aaronson, Barrow, and Sander (2007), using data from Chicago, found that a one standard deviation increase in mathematics teacher effectiveness over a full year raised 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.[2] 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 7 and 10 in Ireland, grades 2 and 5 in Italy, and ages 10 and 12 in Singapore (O'Donnell & Sargent, 2008). Estimating teacher fixed effects models when tests are not administered annually is therefore of considerable policy relevance.

Focusing on teacher effectiveness in the Australian context is also interesting for other reasons. The teaching profession is more regulated than in the United States, with a higher unionization rate, and uniformity in public school teacher salary schedules across entire states and territories (not just across school districts).[3] Moreover, Australia can be regarded as a relatively low-accountability environment, since at the time when the data in this study were collected, test scores were not publicly reported at a school level.

This paper uses data from the state of Queensland, where standardized tests are conducted every two years. With over 90,000 primary school pupils in grades 3–7 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 analyzes the teacher fixed effect terms to see how much of the variation between teachers can be 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 school students between grades 3 and 7 who attended government schools in the state of Queensland during the years 2001–2004.[4] The Queensland Department of Education, Training and the Arts (DETA) administers standardized lit-

[2] 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% of measured variance, while school-level variance constitutes just 4–8% of total variance. A similar study focusing on year 12 students found that class/teacher effects consistently accounted for 59% of the residual variance in student achievement, compared with 5% at the school level (Rowe, 2000; Rowe, Turner, & 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 classroom-level shocks are large.

[3] It is difficult to obtain comparable unionization rates at a fine occupation level, but in 2008 the unionization rate for US workers in 'Education, Training, and Library Occupations' was 43% (Bureau of Labor Statistics), while the unionization rate for Australian 'Education Professionals' was 55% (Australian Bureau of Statistics).

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

Table 1
Cohorts used in the study.

	Cohort 1	Cohort 2	Cohort 3
Year			
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	29,686	30,371	29,745
Numeracy test	29,926	30,588	30,035

Test years marked in italics.

eracy 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.[5]

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 student identifier codes and school codes (plus students' birth dates as a cross-check) 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 student 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.

Because some students move between grades, are absent on the day of the test, or have their birthdates miscoded 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.

The timing of tests in Australia also introduces complications. Previous papers that estimate teacher fixed effect models (such as Rivkin et al., 2005; 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.[6] 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.[7] 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.[8]

The full model, the results of which are shown in columns 3 and 4 of Table 2, is:

$$Y_{ijgt} = T_j + C_{jgt} + \Psi_{gt} + \Pi_i + \varepsilon_{ijgt} \qquad (1)$$

Y_{ijgt} is the literacy or numeracy test score of individual *i*, taught by teacher *j*, in grade *g*, and calendar year *t*, which is modeled as a function of teacher fixed effects T_j, class size C_{jgt}, grade × calendar year fixed effects Ψ_{gt}, student fixed effects Π_i, and a normally distributed error term ε_{ijgt}.[9] In columns 1 and 2 of Table 2, I also show the results omitting

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

[6] Students typically have the same teacher for all subjects. There is no formal system for tracking students by ability across schools at this age. Within schools, principals have discretion over the way in which students and teacher are allocated to classes. In practice, the degree of streaming is relatively small, but the combination of non-random sorting of students across classes and mean-reversion could in principle bias the teacher fixed effects. To address this, I also re-estimate fixed effects at the grade × year × school level, and regress these on the mean characteristics of all teachers in that grade–year–school cell. This strategy produces quite similar results to those shown in Tables 5 and 6.

[7] 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.

[8] 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.

[9] Because I only observe each student and each teacher at a single school, the model does not include school fixed effects.

Table 2
Estimating teacher fixed effects from panel data.

Dependent variable	(1) Literacy test	(2) Numeracy test	(3) Literacy test	(4) Numeracy test
Panel A: dropping non-test years				
F-test for joint significance of teacher fixed effect terms	4.89***	5.82***	2.92***	4.08***
Teacher fixed effects	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	Yes
Class size	Yes	Yes	Yes	Yes
Grade × calendar year fixed effects	Yes	Yes	Yes	Yes
Observations (students × years)	180,116	181,621	180,116	181,621
Number of students	90,840	91,606	90,840	91,606
Number of teachers	9,226	9,233	9,226	9,233
Number of schools	1,057	1,058	1,057	1,058
Panel B: interpolating non-test years				
F-test for joint significance of teacher fixed effect terms	5.48***	6.36***	3.02***	4.19***
Teacher fixed effects	Yes	Yes	Yes	Yes
Student fixed effects	No	No	Yes	Yes
Class size	Yes	Yes	Yes	Yes
Grade × calendar year fixed effects	Yes	Yes	Yes	Yes
Observations (students × years)	268,165	270,391	268,165	270,391
Number of students	90,847	91,613	90,847	91,613
Number of teachers	11,240	11,249	11,240	11,249
Number of schools	1,057	1,058	1,057	1,058

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.

*** Statistical significance at the 1% level.

the student fixed effects. Omitting the class size control or the grade × calendar year fixed effects (or both) has virtually no impact on the results.

An important advantage of this methodology is that the inclusion of student fixed effects means that the results are estimated not from students' performance in a single test, but on the change in student performance over two successive tests. 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.[10]

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 (i.e. students of the average teacher maintain their position in the relative test score distribution).

The results of the student-level regression are shown in Table 2. Columns 1 and 2 show the results without student fixed effects, while columns 3 and 4 show the (preferred) specification with student fixed effects.[11] In each specification, an *F*-test strongly rejects the hypothesis that the teacher fixed effects terms are jointly equal to zero. This is true whether the analysis omits non-test years (Panel A) or linearly interpolates test scores in non-test years (Panel B). In each case, I can easily reject the null hypothesis that there are no systematic differences between teachers.

The coefficients on class size (not reported) tend to be negative for the numeracy tests and positive for the literacy tests. Although the class size coefficients are statistically significant in some specifications, it would be unwise to draw any causal inference from this, given the possibility of non-random sorting of students across differently sized classes.

Not surprisingly, the teacher fixed effects for literacy and numeracy are highly correlated. Estimating the teacher fixed effects using the approach in Table 2, Panel B, columns 3 and 4, and weighting teachers by the inverse of the standard error of their fixed effect, the correlation between teacher fixed effects for literacy and numeracy is 0.37.[12] Fig. 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.

The dispersion of the teacher fixed effects terms provides a measure of the dispersion of teacher performance across Queensland primary schools. However, because

[10] 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.

[11] Computationally, the student fixed effects are estimated by demeaning 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). For Stata users with smaller samples or larger computers, the two-way fixed effects routine outlined in Cornelissen (2008) may be useful.

[12] The standard errors for the teacher fixed effects are estimated in Stata using the *fese* module (Nichols, 2008).

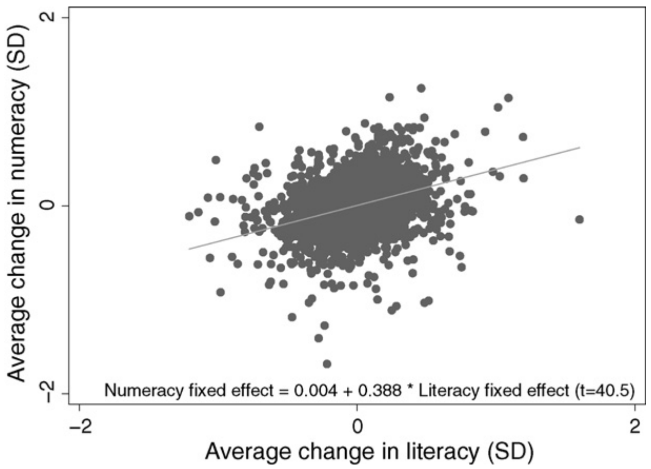

Fig. 1. Are good numeracy teachers also good literacy teachers? *Each dot represents one teacher.*

the teacher fixed effects are measured with error, the observed variance of the teacher fixed effects terms will be larger than the true variance across teachers. As noted in McCaffrey, Lockwood, Koretz, and Hamilton (2003), if we are interested in the optimal estimate for individual teachers, we should use the estimated fixed effects. However, if we are interested in the degree of dispersion, it is necessary to shrink the variance of the fixed effects to account for sampling error.

There are various ways of shrinking the variance.[13] Following Rockoff (2004), I use an empirical Bayes technique employed in the meta-analysis literature, which assumes the teacher fixed effects are normally distributed, and models the observed variance of the fixed effects terms as an additive function of some true variance, plus sampling error. In practice, this is done using the iterative technique outlined in Thompson and Sharp (1999), in which the true variance is estimated as a function of the observed teacher effects and their standard errors.[14] The unadjusted and adjusted standard deviations of the teacher fixed effects terms are set out in Table 3. Using the shrinkage estimator, the standard deviation on the teacher fixed effects terms falls to around 0.13–0.15 when non-test years are dropped, and to around 0.10–0.12 when non-test years are interpolated.[15] 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.

	Unadjusted	Adjusted
Panel A: dropping non-test years		
Literacy	0.194	0.126
Numeracy	0.200	0.146
Panel B: interpolating non-test years		
Literacy	0.156	0.101
Numeracy	0.160	0.116

Note that even adjusting the dispersion for sampling error, there is a wide distribution of teacher fixed effects. Even the most conservative estimate in Table 3 suggests that the adjusted standard deviation of teacher fixed effects is 0.1. This implies 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.

To make the above examples more concrete, note that the most disadvantaged racial group in Australia are Aboriginal and Torres Strait Islander people (Indigenous Australians). In grades 5 and 7, the test score gap between Indigenous students and non-Indigenous students in Queensland is approximately three-quarters of a standard deviation (Bradley, Draca, Green, & Leeves, 2007; see also Leigh & Gong, 2009). This implies that Indigenous students perform approximately 1 1/2 grades below their non-Indigenous counterparts. 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 that the black–white test score gap in Australia could be closed in five years by giving all Indigenous pupils teachers at the 75th percentile.

3. Can teacher demographics explain the variation in teacher effectiveness?

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 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.

[13] Another common technique for shrinking the variance is to estimate separate teacher fixed effects for each year, and extract the persistent portion of each teacher's fixed effect (see, e.g. Kane & Staiger, 2008). However, such an approach involves discarding teachers who are only observed once, which is undesirable in a short panel.

[14] Let t_j be the estimated teacher effect of teacher j, and σ_j be the standard error of that effect. Where $\bar{t}$ is the mean of the teacher effects (zero by construction), and T is the number of teachers in the sample, I estimate the true variance τ^2 as $\tau^2 = \left(\sum w_j^*\right)^{-1} \sum w_j^*(T(T-1)^{-1}(t_j-\bar{t})^2 - \sigma_j^2)$. In the first iteration, $w_j^* = \sigma_j^{-2}$. In subsequent iterations, $w_j^* = (\tau^2 + \sigma_j^2)^{-1}$. The process is repeated until successive iterations of τ^2 differ by less than 0.0001.

[15] Another approach is to simultaneously estimate current teacher and lagged teacher effects. The dispersion in teacher effectiveness with this approach is very similar to that shown in Table 3.

Table 4
Unweighted summary statistics for teacher characteristics.

Variable	Obs	Mean	Std. Dev.
Masters degree or other further qualification	10,398	0.100	0.300
Female	10,398	0.773	0.419
Age	10,398	40.125	10.458
Experience	10,398	13.527	11.055
DETA rating = 1	6,194	0.724	0.447
DETA rating = 2	6,194	0.180	0.384
DETA rating = 3 or 4	6,194	0.097	0.295

Table 4 sets out the characteristics of the 10,398 teachers in the sample (for this part of the paper, I drop teachers with missing demographics or fixed effects). Around 10% have a master's degree or some further qualification.[16] The share of teachers who are female is 77%, the average age is 40, and the average number of years of experience is 14.

The DETA also provides a "suitability rating" for 6194 teachers, or about two-thirds of the sample.[17] This rating is determined by an assessment panel that comprises at least two people, including a teacher and a principal. The panel makes their rating decision based upon an interview, during which the candidate makes a 10-min presentation about their professional experience and their ability to prepare and implement lessons. The interviewee then answers questions for 20–30 min. In the case of applicants who have just completed a teaching practicum at a government school, classroom observation of their performance will also be taken into account. Teachers can request a reassessment, though this may result in a rise or fall in their rating.

Teachers receive a rating of S1 ("outstanding applicants"), S2 ("quality applicants"), S3 ("satisfactory applicants"), or S4 ("eligible for temporary/casual employment").[18] The purpose of the rating is so that teachers can be allocated to positions on merit. Although a rating in the top category is not a prerequisite to teach, the official document describing the ratings process stated that "an S1 applicant will be made an offer of employment before an S2 applicant with similar location preferences and teaching areas" (Education Queensland, 2004, 11). Across the sample, 72% of teachers were rated in the top category, 18% were in the second-highest category, and 10% were 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.

In Table 5, I show the results of regressing the teacher fixed effects on various observable characteristics that are in the DETA payroll database. Panel A shows the results using the teacher fixed effects 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% 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 et al., 2005; Rockoff, 2004), which also find no positive impact of having a master's degree. However, it should be noted that my estimates – and those from the US – are based upon comparing those teachers who chose to obtain master's degrees with those who did not. It is entirely plausible that master's degrees have a positive impact on student test score gains, but that there is some negative selection into master's programs. A preferable estimation strategy would be to observe teachers before and after they obtain a master's degree; but this is not feasible with the present dataset.

There appears to be a statistically 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 in the presence of other controls 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 experience coefficient to become statistically insignificant. In the numeracy specification, the effects of age and experience are larger in magnitude than in the literacy specification, 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 & Ryan, 2008). 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.

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), teachers rated 1 or 2 produce higher test score gains. The positive relationship between the DETA rating and value-added is slightly larger for literacy than

[16] Since the 1980s, registered teachers in Queensland public schools have been required to complete at least four 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 master's degree, a doctorate, or a second degree.

[17] The suitability rating described here was in place in Queensland from 1998 to 2008, after which the four-point scale was replaced with a six-point scale (presumably in an effort to achieve greater dispersion of ratings).

[18] 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.

Table 5
Student test score gain and teacher characteristics.

	(1)	(2)	(3)	(4)	(5)
Panel A: literacy					
Masters	−0.0101** [0.0040]				−0.0079* [0.0040]
Female		0.0140*** [0.0026]			0.0153*** [0.0027]
Age			0.0004*** [0.0001]		0.0003* [0.0002]
Experience				0.0004*** [0.0001]	0.0002 [0.0002]
R-Squared	0.0006	0.0022	0.0012	0.001	0.0043
Teachers	10,398	10,398	10,398	10,398	10,398
Panel B: numeracy					
Masters	−0.0125*** [0.0040]				−0.0083** [0.0041]
Female		−0.0050* [0.0028]			−0.0021 [0.0029]
Age			0.0009*** [0.0001]		0.0004** [0.0002]
Experience				0.0009*** [0.0001]	0.0005*** [0.0002]
R-Squared	0.0008	0.0003	0.0049	0.0055	0.0064
Teachers	10,398	10,398	10,398	10,398	10,398

Dependent variable is the teacher fixed effect. 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 inverse of the standard error on each teacher's fixed effect.

* Statistical significance at the 10% level.

** Statistical significance at the 5% level.

*** Statistical significance at the 1% level.

for numeracy. Controlling for gender, qualifications, age, and experience, the relationship between value-added and the top DETA ranking is statistically significant at the 1% level for literacy and the 10% level for numeracy.

Since Tables 5 and 6 only include experience as a linear term, Figs. 2 and 3 test whether there is a nonlinear relationship between experience and student test score gain. Both charts are based upon a kernel-weighted local polynomial regression of teacher fixed effects on experience, with a gray band depicting the 95% confidence interval.[19]

Table 6
Student test score gain and education department ratings.

	(1)	(2)
Panel A: literacy		
Rating = 1	0.0087* [0.0051]	0.0149*** [0.0053]
Rating = 2	0.0057 [0.0061]	0.0104* [0.0062]
Masters		−0.0051 [0.0047]
Female		0.0078** [0.0039]
Age		0.0001 [0.0002]
Experience		0.0007*** [0.0003]
R-Squared	0.0004	0.0035
Teachers	6194	6194
Panel B: numeracy		
Rating = 1	0.0023 [0.0056]	0.0114* [0.0060]
Rating = 2	0.0055 [0.0067]	0.0111 [0.0068]
Masters		−0.0032 [0.0049]
Female		−0.0078* [0.0040]
Age		0.0004** [0.0002]
Experience		0.0008*** [0.0003]
R-Squared	0.0001	0.0048
Teachers	6194	6194

Dependent variable is the teacher fixed effect. 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 inverse of the standard error on each teacher's fixed effect. The DETA rating ranges from 1 to 4, with 3 and 4 being the excluded category from the regressions (only one teacher in the sample is rated 4).

* Statistical significance at the 10% level.

** Statistical significance at the 5% level.

*** Statistical significance at the 1% level.

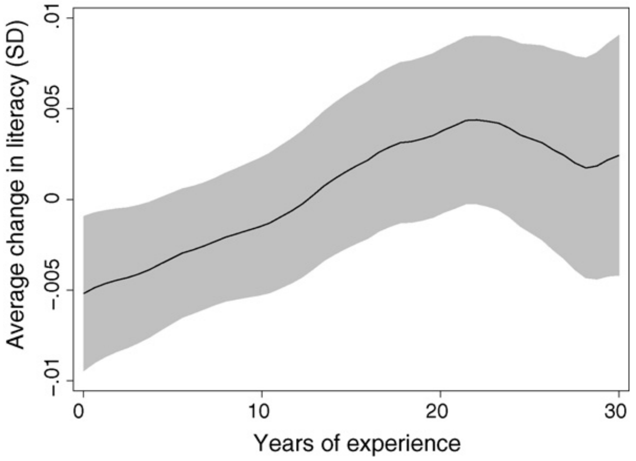

Fig. 2. Literacy and teacher experience.

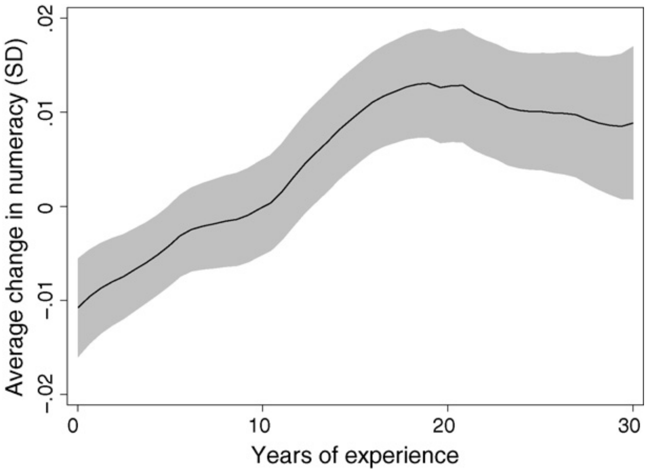

Fig. 3. Numeracy and teacher experience.

For both literacy and numeracy, there appears to be a statistically significant effect of experience in the early years. Compared to novice teachers, teachers with twenty years of experience have test score gains that are 0.1 standard deviations higher in literacy, and 0.2 standard deviations higher in numeracy. Beyond twenty years, there appear to be no further gains to experience (indeed, there is some suggestion of a drop in value-added for literacy, but this is not statistically significant).

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.

Unfortunately, while it is possible to draw conclusions about the differences in effectiveness between teachers, there is little evidence on the cost of policies to improve teacher quality (Hanushek & Rivkin, 2006). However, it is possible that raising teacher quality may be at least as cost-effective 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.[20] The comparison would favor teacher quality still more if the benefits of class size reductions are smaller than the estimates of Krueger (1999) (e.g. Hanushek, 1998; Hoxby, 2000 find zero or negligible benefits of across-the-board class size reductions); or if large-scale class size reductions have the effect of lowering teacher quality in disadvantaged schools (see, e.g. Jepsen & Rivkin, 2009).

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 (with the experience gradient being steeper for numeracy). 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% of the variation in teacher performance remains unexplained by differences in measured teacher demographics. This suggests that uniform pay schedules are 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.

Acknowledgments

The paper has been improved by comments and suggestions from Associate Editor Mikael Lindahl, an anonymous referee, 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. Susanne Schmidt and Elena Varganova provided outstanding research assistance. The support of the Australian Research Council's Discovery funding scheme (project number DP0665260) is acknowledged. While all care has been taken in preparing this publication, the State of Queensland (acting through the Department of Education, Training and the Arts) does not warrant that the publication is complete, accurate, or current. The Department of Education, Training and the Arts expressly disclaims any liability for any damage resulting from the use of the material con-

[19] I use an epanechnikov kernel, with a polynomial of degree zero (i.e. local-mean smoothing), and the default bandwidth (3.6 for literacy and 2.4 for numeracy). As with the results in Tables 5 and 6, estimates are weighted by the inverse of the standard error on each teacher's fixed effect.

[20] For example, using panel data on starting teacher salaries across Australian states and territories, Leigh (2009) finds that a 1% rise in the salary of a starting teacher boosts the average aptitude of the future teaching pool (students entering teacher education courses) by 0.6 percentile ranks.

tained in this publication and will not be responsible for any loss, howsoever arising from use of, or reliance on this material.

References

Aaronson, D., Barrow, L., & 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., & Margolis, D. N. (1999). High wage workers and high wage firms. *Econometrica*, *67*(2), 251–333.

Bishop, J. (1991). Achievement, test scores and relative wages. In M. H. Kosters (Ed.), *Workers and their wages* (pp. 146–186). Washington, DC: AEI Press.

Bradley, S., Draca, M., Green, C., & Leeves, G. (2007). The magnitude of educational disadvantage of indigenous minority groups in Australia. *Journal of Population Economics*, *20*, 547–569.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., et al. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cornelissen, T. (2008). The Stata command felsdvreg to fit a linear model with two high-dimensional fixed effects. *Stata Journal*, *8*(2), 170–189.

Currie, J., & Thomas, D. (2001). Early test scores, socioeconomic status, school quality and future outcomes. *Research in Labor Economics*, *20*, 103–132.

Education Queensland. (2004). *Applying for teacher employment booklet 2004*. Brisbane: Queensland Government.

Hanushek, E. A. (1998). *The evidence on class size*. W. Allen Wallis Institute of Political Economy Occasional Paper Number 98-1. Rochester: University of Rochester.

Hanushek, E. A., & Raymond, M. E. (2002). Improving educational quality: How best to evaluate our schools? In *Paper prepared for education in the 21st century: Meeting the challenges of a changing world* Federal Reserve Bank of Boston, June 19–21, 2002.

Hanushek, E. A., & Rivkin, S. G. (2006). Teacher quality. In E. A. Hanushek, & F. Welch (Eds.), *Handbook of the economics of education* (pp. 1051–1078). Amsterdam: Elsevier.

Hill, P. W., & 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., & Rivkin, S. G. (2009). Class size reduction and student achievement: The potential tradeoff between teacher quality and class size. *Journal of Human Resources*, *44*(1), 223–250.

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation*. NBER Working Paper No. 14607. Cambridge, MA: National Bureau of Economic Research.

Krueger, A. (1999). Experimental evidence of education production functions. *Quarterly Journal of Economics*, *114*(2), 497–532.

Leigh, A. (2009). *Teacher pay and teacher aptitude*. Canberra: Australian National University., mimeo.

Leigh, A., & Gong, X. (2009). Estimating cognitive gaps between indigenous and non-indigenous Australians. *Education Economics*, *17*(2), 239–261.

Leigh, A., & Ryan, C. (2008). How and why has teacher quality changed in Australia? *Australian Economic Review*, *41*(2), 141–159.

Marks, G., & Fleming, N. (1998a). *Factors influencing youth unemployment in Australia 1980–1994*. Longitudinal Surveys of Australian Youth Research Report No. 7. Melbourne: ACER.

Marks, G., & 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.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: Rand Corporation.

Murnane, R. J., Willet, J. B., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics*, *77*, 251–266.

Nichols, A. (2008). *fese: Stata module calculating standard errors for fixed effects*. , available at http://ideas.repec.org/c/boc/bocode/s456914.html

O'Donnell, S., & Sargent, C. (2008). *International review of curriculum and assessment frameworks. INCA comparative tables* (July 2008 ed.). London: Qualifications and Curriculum Authority and National Foundation for Educational Research., available at http://www.inca.org.uk/

Rivkin, S. G., Hanushek, E. A., & 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.

Rowe, K. (2000). Celebrating coeducation? Maybe, but not necessarily for academic achievement! An examination of the emergent research evidence. In *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., & Lane, K. (2002). Performance feedback to schools of students' year 12 assessments: The VCE data project. In A. J. Visscher, & R. Coe (Eds.), *School improvement through performance feedback* (pp. 163–190). Lisse, The Netherlands: Swetz & Zeitlinger.

Rowe, K. J., Turner, R., & 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. In *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 to December 2, 1999.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*(20), 2693–2708.