ECONOMIC RECORD

ECONOMIC RECORD, VOL. 94, SPECIAL ISSUE, JUNE, 2018, 73–101

Do School Principals Respond to Increased Public Scrutiny? New Survey Evidence from Australia*

MICHAEL COELLI
Economics, University of Melbourne, Melbourne, VIC, Australia

GIGI FOSTER
Economics, University of NSW, Sydney, NSW, Australia

ANDREW LEIGH
Parliament of Australia, Canberra, ACT, Australia, Institute of Labor Economics, Bonn, Germany

We explore responses of Australian school principals to the introduction of test score reporting via the My School website in 2010. Theory suggests that heightened public scrutiny should motivate principals to adopt best practices for improving their schools' test results. We use responses from both public and private schools to a custom-built questionnaire administered to principals before (2009) and after (2012) the My School website launch. We find scarce evidence of meaningful adjustments over time, but we do find evidence of significantly different policies and practices across school groups.

I *Introduction*

The existing literature on school accountability suggests that schools improve their performance on standardised exams when they are held accountable for this performance and when performance information is publicly accessible.[1] These studies suggest that the improvements in test scores associated with school accountability are large in comparison to those attainable via many other types of educational interventions (Lee, 2008). Evidence exists that schools facing accountability pressure respond to this pressure in substantive ways, although the literature also offers many examples of ways in which schools respond to accountability pressures to affect measured performance without contributing to generalised improvements in outcomes.[2] Most studies – with a few recent exceptions, such as

*This project would not have been undertaken without David Figlio, who helped initiate the broad project that underpins this paper, including design and implementation of the 2009 survey. Our thanks go to Van Dao, Narin Chhun and the rest of the ACARA team for their support in extracting and explaining the NAPLAN data, and Kenneth Mease for advice on survey design. We also thank Gordon Leslie, Su-Min Lim, and Joseph Madden for superior research assistance, and the University of New South Wales and the University of Melbourne for financial support. All errors remain ours and all views expressed herein are ours, and not those of ACARA or any other institution.

JEL classifications: I21, I28

Correspondence: Michael Coelli, Economics, University of Melbourne, Melbourne, VIC, Australia. Email: mcoelli@unimelb.edu.au

[1] Figlio and Ladd (2015) and Figlio and Loeb (2010) provide surveys of this large literature. Some standalone analyses include Carnoy and Loeb (2002), Chiang (2009), Dee and Jacob (2011), Deming *et al.* (2016), Figlio and Rouse (2006), Hanushek and Raymond (2004), Hussain (2013), Neal and Schanzenbach (2010), Reback *et al.* (2014), Rockoff and Turner (2010), Rouse *et al.* (2013) and West and Peterson (2006).

[2] For instance, numerous studies (e.g. Haney, 2000; Booher-Jennings, 2005; Neal & Schanzenbach, 2010) show that schools subject to accountability pressure concentrate their energies on high-stakes rather than low-stakes subjects, teach skills that are valuable for the specific tests with which they are assessed rather than other potentially important skills, and emphasise the education of students most likely to contribute either positively or negatively to the school's rating. Other research demonstrates that schools facing accountability pressure attempt to affect outcomes through exclusions of selected students from testing (Cullen & Reback, 2006; Figlio & Getzler, 2006; Coelli & Foster, 2016), selective discipline (Figlio, 2006), changing school meal plans on the day of the test (Figlio & Winicki, 2005) or even outright cheating (Jacob & Levitt, 2003).

doi: 10.1111/1475-4932.12400

Mizala and Urquiola (2013) and Andrabi *et al.* (2017) – focus on the impact of accountability only on public schools' responses, due to the inaccessibility of data on private schools.

A small and very recent literature (Coelli & Foster, 2016) has exploited the availability of data on all Australian school sectors – public, independent and Catholic – to examine accountability effects on Australian schools resulting from the launch of a website called My School. The Australian government introduced this website in 2010 to increase public scrutiny of school performance by publicly disseminating multidimensional results on the National Assessment Program – Literacy and Numeracy (NAPLAN) national tests. The data used in these recent papers are drawn from the Australian Curriculum and Reporting Authority (ACARA), which collects NAPLAN scores for the country and shares them with researchers at the level of the school or the student on a case-by-case basis, as detailed in Pugh and Foster (2014).

Most studies of school accountability effects treat the mechanism by which such effects materialise, particularly on the supply side, as a black box. What exactly do schools do to substantively improve student performance? Rouse *et al.* (2013) provide unique insight into this question by surveying Florida public school principals at *ex post* high-performing and low-performing schools before and after an accountability intervention that involved the release of updated school performance indicators. Data from their surveys, consisting of scientifically developed survey items that ask about a range of school policies and practices, indicate that at the time of the first survey, many poorly performing schools were already engaging in a range of interventions generally thought to be associated with good educational outcomes. Comparing data across survey years, they find that relative to other schools, the worst-graded schools change their policies and practices to give more attention to low-performing students, increase instruction time, and increase the flexibility and/or generosity of the scheduling, resourcing, and/or decision-making environment facing teachers.

In this paper, we use new survey data from Australian school principals to analyse the impact of the major information shock represented by the launch of My School in 2010. In the months before the website went live, we set about surveying as many Australian school principals as possible, asking them a detailed battery of questions about everything from budget autonomy to homework requirements. Three years later, we surveyed principals from responding schools to see how their schools had changed in the wake of test score reporting.

In addition to the magnitude of the information change represented by My School's launch, another strength of our study over prior research is the breadth of the schools in the sample. Our sample includes primary and secondary schools in the public, independent and Catholic sectors. This diversity allows us to test for differential responses to Australia's uniquely sudden and significant increase in school accountability.

At baseline, we find striking differences in the initial policy settings in place at low-performing and high-performing schools. At low-performing schools, we find that parents are less involved, teachers have lower expectations of students and spend less time with students outside the classroom, fewer hours are assigned to teachers for planning and reviewing, minimum class time for several subjects is less likely to be mandated, and the school day is shorter, although regular classes are smaller and teacher assistance and tutoring of low-performing students are more likely to be used in the classroom. We also find significant differences in initial policies and practices across the three Australian school sectors, with independent schools, for example, allocating the most time across all three sectors for teacher preparation, being most likely to provide tutoring outside of class and to set reduced class sizes for gifted students, being the likeliest to feature incentives for teachers (including dismissal), and requiring the lengthiest homework commitments by students in mathematics and reading.

Comparing principals' responses before and after the launch of My School, we find little evidence that low-performing schools respond to the accountability shock in terms of their overall or student-focused policies and practices. Low-performing schools in fact fall even further behind other schools in terms of setting minimum class time and time assigned for teacher preparation, although they do increase even further their already relatively high use of classroom-based assistance for teachers (while in non-government low-performing schools, the use of teacher assessment also rises). There is also some evidence of responses by low-performing schools targeted to specific subjects, both in terms of overall curriculum narrowing and, in low-performing government primary schools, the redirection of resources towards the curriculum area (whether literacy or numeracy) in which a school performed poorly and away from the other area.

In summary, our evidence shows that poorly performing schools in Australia feature policies and practices that the education literature generally deems worse for students, and further that low-performing schools in general do not react to the publication of their poor performance on My School by substantially changing their overall or student-focused policies and practices. We conclude by briefly evaluating possible reasons for our results, including that principals are not focused on optimising their schools' published performance, that rigidities in the education policy-setting environment prevent principals from adjusting the way their schools are run, and that the dramatic move from a low-information to a high-information environment takes more than a few years to affect how schools operate.

The remainder of the paper is organised as follows. Our empirical approach is described and justified in Section II, which also includes details of the launch of the My School website and of our surveys of school principals. Our main estimates and results are presented and discussed in Section III, and we offer some concluding remarks in Section IV.

II Empirical Approach

We survey principals of Australian schools of all levels (elementary, secondary and combined) and all sectors (public, independent and Catholic) before and after a simple accountability shock, asking about the policies and practices in place at their school. We first document differences by NAPLAN performance and school sector in principals' responses to our first survey, and we then apply a simple differencing method to examine the changes in policies and practices that principals report over time.

If school principals care about the increase in access to knowledge about their schools' NAPLAN performance that My School provides, presumably because they face monetary or non-monetary incentives that relate either directly or indirectly to this increase, then the principals of more poorly performing schools should be expected to try to improve their schools' performance on the NAPLAN tests once My School is launched, possibly by adjusting the policies and practices in place at their schools. As we have no convincing evidence on the strength or otherwise of such incentives, we take a revealed-preference approach to this question: if our statistical evidence indicates that the policies and practices in place at poorly performing schools do not change in response to My School, then one possible reason for this is that Australian principals are not incentivised to try to improve their schools' performance.

(i) The Accountability Intervention

In the prior study most similar to ours, Rouse *et al.* (2013) evaluate the impact of a complicated accountability intervention, where school performance grades issued in summer 2001, under the auspices of a state-wide accountability policy first introduced in 1999, were updated in June 2002. As a result of this update, some schools may have been caught by surprise with their new (different) grade. The authors measure changes in policies and practices between spring 2002 and spring 2004 for all schools receiving fail-level grades in 2002, not only for those whose grades changed. Many schools retained the same grade before and after the 2002 regrading, however, arguably diluting the estimated effects of the accountability shock.

While the Florida experiment represented a change to the available information about school performance, the Australian experiment represented a massive increase in the amount of information that parents had about their school's performance. As the Organisation for Economic Co-operation and Development (OECD) notes: 'Prior to the advent of My School, parents of school children were unable to understand the operations and achievements of their schools on common national definitions and measures' (OECD, 2012, p. 9). The My School website launch on 28 January 2010 that we study was the first time that comprehensive absolute and relative academic performance data were publicly disseminated to all schools via a single access point. The OECD compares the demand to access the My School website on launch day to the demand for large news sites and popular reality television shows in an entire month (OECD, 2012, p. 35). On the day after the launch, school results were discussed on the front pages of all major Australian newspapers. On average, the My School website received around 8,000 unique visitors per day during 2010, and over 2,000 unique daily visitors in subsequent years.[3]

It is difficult to overstate the scale of the information shock that the My School website represented. Prior to the My School launch, most schools did not publish any information about their test score performance. In certain instances, information about school performance was available through informal networks or in schools' annual reports, but it was rarely possible to compare schools' performance.[4] Reflecting the scale of the proposed change, test score reporting was opposed by the Australian Secondary Principals Association, the Australian Parents Council, the Independent Education Union of Australia, the Australian Government Primary Principals Association, the Australian Education Union and the Australian Council of State School Organisations (Patty, 2009). Industrial action was threatened if it went ahead.

To what extent was the information published on the My School website new to school principals? While Australian states had run literacy and numeracy testing programs for many years, nationally comparable testing across all states using the NAPLAN instrument only began in May 2008, and the first results of this testing were only published internally to schools in September 2008. Even if the information schools received in September 2008 had been an exact replica of what was eventually published publicly on My School (which is highly doubtful), principals would have had less than a year to take remedial action before our first survey was put into the field, and may not have felt strongly pressured to do so in any event, given that the 2008 performance data had not yet been made public. In addition, our measure of school performance combines schools' scores from 2008 and 2009, the latter of which were only known to schools in September 2009, around the time our first survey was put into the field. We amalgamate test score information for these two years because both sets of scores were published on the My School website simultaneously upon website launch.

(ii) School Performance Measurement

In the typical setting examined in prior research, schools are divided starkly into 'winner' and 'loser' groups. This is true in the case of the school awards examined by Mizala and Urquiola (2013) and in Rouse *et al.* (2013), whose performance data take the form of a letter grade, where 'F' denotes failure and signals the potential for intervention. By contrast, the school performance signals sent via My School are more or less continuous, as they are in the form of average national test scores across students in tested cohorts at the school. While My School also provides colour coding, including red for 'worse' than other schools, there is no threshold below which a school is labelled as having 'failed'. Similarly, although green coding is used to denote 'better' performance than other schools, there is no threshold above which an Australian school receives a prize or is seen to have decisively won a performance contest.

Appendix II provides two screenshots from My School:[5] one for a relatively high-performing primary school, and one for a relatively low-performing high school. On each screenshot, the average NAPLAN test scores of an individual school are shown together with the relevant colour coding. NAPLAN assesses students on five different domains of learning (numeracy, reading, grammar/punctuation, writing and spelling), and is administered to children in school grades 3, 5, 7 and 9 in every year. The website displays all available historical score averages by domain and grade for each school. Test score comparisons are shown in the form of numbers and colours against both 'similar' schools offering the tested grade (denoted 'SIM' in the screenshot) and against all schools in Australia that offered the tested grade (denoted 'ALL'). The 'similar' school comparisons are based on the average test score across up to 60 schools identified by ACARA as being similar to the

[3] Specifically, the average number of unique daily visitors was 7,976 in 2010, 2,700 in 2011, 2,390 in 2012, 2,101 in 2013, 2,700 in 2014, 2,376 in 2015, 2,221 in 2016 and 3,510 in 2017 (figures provided by ACARA).

[4] Prior to the My School launch, school-level test scores were reported in Tasmania and Western Australia, though the Western Australian data were only presented in graphical form. In addition, some states reported school-level grade-12 results. At a local level, schools were permitted to report their results in their annual reports, but few did so, and these data were not compiled in any comparable form for parents and other stakeholders.

[5] These screenshots represent the material as presented at the time of our surveys. My School has updated the look of its website for the release of the 2017 NAPLAN data, but the same colours are used.

focal school based on parental occupation and education,[6] remoteness of the school, and Indigenous student percentage.[7] Each coloured indicator that accompanies a school's average score for each domain and grade within a cohort denotes the level of the school's performance relative to similar and all schools, respectively. The colour of each indicator is determined by the distance of the school's score in that domain × grade × cohort to the mean of the relevant comparison group (either similar schools or all schools). Dark green (dark red) colouration is used to denote a score that is 'substantially' higher (lower) than the comparison-group mean. Light green (light red) denotes a score that is 'somewhat' higher (lower) than the comparison-group mean. Absence of colour denotes a score that is approximately the same as the comparison-group mean. The thresholds that define 'substantially' and 'somewhat' are respectively 0.5 times and 0.2 times the standard deviation across all students in Australia of scores in that domain × grade × cohort. Thresholds are calculated using the standard deviation across all school students in Australia, even when the resulting colouration reflects a comparison to the similar-school mean.

With five learning domains and potentially four grade cohorts assessed by NAPLAN each year, and two comparisons for each of these domain × grade × cohort cells shown on My School for a given year – one comparing the focal school to all Australian schools, and the other comparing the focal school to similar schools – there are many possible ways to reduce the information displayed on My School into a single metric.

For analytical tractability, we reduce the multidimensionality of this information by first calculating the percentage, across all tested grades in the school and across all five learning domains, of a school's scores that fall into each colouration category (dark green, light green, clear, light red, dark red). For example, if a school offered only grades 7–12 (a normal 'high school' in most Australian states),[8] then to calculate the 'per cent dark red' measure for this school for a given year, we would: (i) calculate the percentage of the average scores posted by the school that year for tested grade 7 students across all five learning domains that fell into the 'dark red' zone; (ii) calculate the analogous percentage for tested grade 9 students; and (iii) take the simple average of these two percentages. We then use the 'per cent [colour]' variables – primarily the 'per cent dark red' variable – constructed in this fashion to build dummy variables that indicate performance bands. We choose this method of constructing a performance measure with the aim of recovering a value that is as close as possible to the rough 'feel' about a school's performance that a parent or other stakeholder would get while browsing through the school's posted results on My School.

(iii) Surveys

In 2009 we sent invitations to principals of all 9,552 Australian schools to complete our initial survey, and then in 2012 we sent invitations to complete our second survey to principals of the 1,929 schools whose principals had responded to our 2009 survey and which were still operating as separate entities in 2012. Appendix III contains further details regarding the implementation of the surveys. With data from the surveys in these two years, we can examine both how schools' initial policies and practices differ across school groups, and how schools' policies and practices changed between one year before and two years after the My School launch. In the absence of an additional pre-My School survey, we are unable to control for any pre-My School trends in policies and practices.

Responses to our first school principal survey were received between 11 October 2009 and 29 January 2010, prior to the start of the 2010 school

[6] In 2008 and 2009, direct reports of parental education and occupation were not collected. Measures of average education and occupation in the post code where the family lived were used as proxies in those years.

[7] 'Similar' schools were identified using ACARA's Index of Community Socio-Educational Advantage (ICSEA). The three components of parental occupation and education, remoteness, and Indigenous student percentage are combined to form the ICSEA score for each school in a manner that best predicts student test scores (ACARA, 2014).

[8] Australian primary schools generally cover grades from a pre-grade-1 year (kindergarten or preparatory grade) to grade 6 or grade 7, depending on the state. High schools then cover from grade 7 or 8 to grade 12. In some cases, the high school grades may be split into a junior high school for grades 7–10, and a senior high school or 'college' for grades 11 and 12. The number of grades given the NAPLAN test at a school ranges from one grade, in some high schools that offer education for grades 8–12 only, up to four grades in schools that offer education over all grades from grade 1 or below up to grade 12, referred to as 'combined schools'.

year. Our response rate for this survey was approximately 21 per cent, which we regard as a reasonable response rate for a survey of busy professional leaders that took up to 25 minutes to complete. Of the 1,929 school principals sent invitations to complete the second survey in 2012, approximately 58 per cent responded. These responses provide us with information for both years on 1,122 schools. When conducting our analyses, we restrict our attention to the 1,062 of these 1,122 schools that are classified as standard schools.[9]

In Tables A1 and A2 we present some evidence about the selectivity of the samples of standard schools that responded to our two surveys. For the initial survey we look at selection relative to the whole Australian standard school population (Table A1); and for the second survey we look at selection relative to those standard schools from which responses were received to the 2009 survey (Table A2).

Table A1 shows that the characteristics of schools that responded to our 2009 survey were very similar to the characteristics of those that did not respond. Respondent schools were distributed similarly to non-respondent schools across sector (government, independent, Catholic), location (metropolitan, provincial, remote, very remote), and type (primary, secondary, combined). Respondent schools were also of similar size, had similar fractions of students with a language background other than English (LBOTE), and reported similar average normalised NAPLAN test scores.[10]

The only characteristics on which statistically significant differences by respondent status are evident in 2009 are the percentage of students from an Indigenous background, which was 1.6 percentage points lower among respondent schools than among non-respondent schools, and ICSEA (an amalgamation of parental background, remoteness and Indigenous background), which is slightly higher (equivalent to 0.05 standard deviations) among respondent schools. After controlling for state by sector by type by location fixed effects, however, differences in Indigenous background and ICSEA are much smaller and for ICSEA no longer statistically significant, as shown in the last two columns of Table A1.

As detailed in Appendix III, substantial effort was undertaken to increase response rates. For example, in the 2012 survey, we offered a prize and an extension of response time to principals requiring it, sent up to four separate letter/email invitations, and finally called schools individually. Table A2 provides summary statistics for schools that did and did not respond to the 2012 survey, treating as the base population those schools that responded to the 2009 survey. In this case, significant differences are observed by respondent status. Respondent schools were larger, had higher ICSEA, had a lower Indigenous percentage and a higher LBOTE percentage, were more likely to be from a metropolitan than a provincial area, were more likely to be Catholic than government, were more likely to be secondary schools than primary schools, and had slightly higher average normalised scores. The lower response rate among government provincial primary schools is due in part to difficulties in obtaining responses from Queensland government schools, driven by the fact that our initial survey web link at first invitation was inaccessible via Queensland government computers (a problem that was fixed once we identified it). Once we control for state by sector by type by location fixed effects, however, there are no statistically significant differences in schools' size, ICSEA, Indigenous percentage, LBOTE percentage or average normalised scores by respondent status, as shown in the last two columns of Table A2.

We implemented our surveys in the form of a core module plus one of four additional modules for each principal. The full set of survey instruments used in 2009 and 2012 is provided in Coelli *et al.* (2018).

[9] Although we also surveyed special schools, which provide education solely for students with learning and other disabilities, such schools are unlikely to respond to the My School website as test score information on such schools is not provided on My School, and most of the students at such schools are not required to sit NAPLAN tests.

[10] Note that the statistics provided in Tables A1 and A2 are not weighted by school size. The slightly negative means of the averaged normalised scores observed for both groups are due to the normalisations being constructed using the means and standard deviations in test scores constructed using all individual Australian students. Generally, students in larger schools perform better on NAPLAN tests. In the unweighted means provided in Tables A1 and A2, the higher test scores among students in larger schools are essentially being underweighted at the student level. These average normalised scores have cross-school means much closer to zero when calculated using weights based on the number of students in a school sitting the NAPLAN tests.

(iv) Measures of School Policies and Practices

The main advantage of our survey data is the breadth of information on how schools are run. We have responses to over 60 separate questions on policies and practices used in schools that might theoretically affect NAPLAN scores, with sub-questions within several main questions. The challenge is to distil that information into tractable measures of school policies and practices that may influence NAPLAN scores.[11] To reduce the dimensionality of the estimation problem, we group responses to individual questions into 16 conceptual 'spheres',[12] namely, *low-performing students*, *lengthening instructional time*, *reduced class size for subject*, *narrowing of curriculum*, *low-performing teachers*, *teacher assigned time*, *school climate*, *control (teacher, state and principal)*, *reduced class size for gifted students*, *teacher time spent outside school hours*, *teacher observed in the classroom*, *assistance in the classroom*, *homework time expected for tested subjects*, *teacher incentives*, *assessment of teachers* and *teacher dismissal frequency*.[13]

After grouping questions into these spheres, we construct measures and corresponding estimates for each sphere. We first map schools' responses about each policy or practice that might theoretically increase NAPLAN scores into a range that flows logically from 'low' to 'high', where low values mean little of the policy or practice is in effect at the school, and high values indicate that it is strongly in effect. We then normalise these individual responses to have mean 0 and standard deviation 1 across all schools, and construct the simple average for each school of its normalised responses within each sphere. We use the mean and standard deviation of responses across responding schools in 2009 to construct the normalisations in both 2009 and 2012. This allows policies and practices as we measure them across all schools to change in aggregate in response to the My School accountability shock.

As noted above, response rates to our surveys, particularly in 2012, differ by state, location, sector and type. To address the bias this might otherwise cause, we employ weights when constructing all our measures and estimates that take account of the differential response rates by state–location–sector–type cells. These weights were constructed as the inverse of the probability of responding to our surveys by state–location–sector–type cells.[14]

(v) Modelling Approach

We construct our main estimates of the effect of revealed school performance on the My School website on school policies and practices by first estimating models at the individual school level for each individual policy or practice P_{jt} implemented in 2012, as follows:

$$P_{jt} = \alpha_j + \beta_j^1 \cdot \mathrm{DR1}_{t-1} + \beta_j^2 \cdot \mathrm{DR2}_{t-1} + X_{t-1} \cdot \Gamma_j + \varepsilon_{jt} \tag{1}$$

Here, $\mathrm{DR1}_{t-1}$ and $\mathrm{DR2}_{t-1}$ are indicators of relatively poor school test score performance as first revealed on the My School website in 2010 (based on the percentage of the school's NAPLAN test scores in 2008 and 2009 identified with dark red flags) and X_{t-1} is a vector of school-level variables: (i) the relevant fully interacted set of state × location × sector × type indicators; (ii) the 2009 levels of school enrolment, ICSEA, Indigenous percentage and LBOTE percentage (for 2010); as well as, importantly, (iii) the school's measure of the policy or practice P_j in 2009. The indicator $\mathrm{DR1}_{t-1}$ denotes a school whose proportion of test scores in 2008 and 2009 with dark red flags lies above 0 but below 0.2, while $\mathrm{DR2}_{t-1}$ denotes a school whose proportion of dark red flags is 0.2 or higher. When we implement this estimation approach, we use the raw measures of the policies and practices, rather than normalised ones.

The β_j^1 and β_j^2 coefficients are our objects of interest, as they indicate whether the implementation of the specific policy or practice in 2012 was correlated with a school's reported poor relative performance on My School in 2010, even

[11] This reduction in dimensionality also assists in minimising the well-known problem of the build-up of Type I errors when conducting multiple tests at once, for which a variety of corrections have been proposed (e.g. Lix & Sajobi, 2010).

[12] Due to the similarity in our questionnaires, many of the spheres we look at correspond to analogous domains in Rouse *et al.* (2013).

[13] The questions on teacher incentives, assessment of teachers, and teacher dismissal frequency were only asked of principals at non-government schools in our survey, as such interventions were difficult if not impossible for local leadership to provide in Australian government schools at the time of the surveys.

[14] A discussion of inverse probability weighting can be found in Hogan and Lancaster (2004).

controlling for the degree of implementation of that policy or practice at the same school in 2009. We standardise each β_j^i estimate by dividing by the standard deviation (σ_j) of schools' responses in 2009 regarding the implementation of the given policy or practice. Our estimates of the effect of the public revelation of school performance on each policy sphere are then the average of the J standardised βs within each sphere d:

$$\beta_d^i = \frac{1}{J}\sum_{j=1}^{J}\frac{\beta_j^i}{\sigma_j} \qquad (2)$$

To obtain the correct standard error of the β_d^i, we take account of potential covariances among the estimates of the various β_j^i within sphere d. To do this, we follow Kling and Liebman (2004) and estimate seemingly unrelated regressions (SURs) using Equation (1) for all policies within each sphere, recovering identical coefficient estimates as obtained under ordinary least squares. We then calculate the standard error of each β_d^i in Equation (2) using the full variance–covariance matrix we construct for the SUR model.[15] A potential advantage of these sphere-level estimates is that while estimates of each β_j^i may be statistically insignificant, estimates of the β_d^i may be significant due to covariation among the outcomes.

III Results

(i) School Policies and Practices Prior to My School

We begin by exploring the initial policies and practices employed by schools prior to My School, as revealed in responses to our 2009 survey. We first focus on government primary schools, as this is the largest homogeneous group of schools in our sample, permitting a relatively clean insight into differences by initial school performance. In Table 1, schools' responses are tabulated separately based on their relative performance in the 2008 and 2009 NAPLAN tests. We separate schools into three performance groups: 'poorly performing' schools (in which 20% or more of the reported domain × grade × cohort NAPLAN test scores were accompanied by dark red flags), 'underperforming' schools (with between 0 and 20% of scores accompanied by dark red flags) and remaining schools (no dark red flags). To create these groups, we use the colouration flags pertaining to the test score comparisons with all schools rather than similar schools. Of our sample of government primary schools, 45 per cent had no dark red flags, while 35 per cent of schools had 20% or more dark red flags.

Table 1 reveals substantial differences between poorly performing schools and schools receiving no dark red flags at all. In poorly performing schools, parents are less involved, teachers have lower expectations of students, fewer hours are assigned to teachers for planning and reviewing, minimum class time for mathematics, reading, writing and art is less likely to be mandated, and the school day is shorter. Teachers in these schools who are judged by their principal to need assistance are more likely to have a teacher's aide assigned to them and more likely to be assigned to coaching directly by the principal, but less likely to have a mentor or lead teacher assigned to help. Poorly performing schools are also less likely to have reduced class sizes for at least one subject to cater for students with academic difficulties or those for whom English is a second language, although they are more likely to have used reduced class size to teach the basic subjects of reading and writing to regular students. Teachers in poorly performing schools spend less time on music, sport, tutoring and field trips, and parents in such schools are less likely to be required to sign their child's homework. In the *control* sphere, principals at poorly performing schools report less teacher control of curriculum and hiring, less principal control of curriculum, hiring and budget, and more state control of everything, including all the above and teacher evaluations. These results are consistent with an overall picture in which poorly performing schools are struggling to serve a disadvantaged population and at which fewer levers of local discretion appear to be available.

Table A3 shows that schools in the two lower-performing groups are smaller and have lower ICSEA scores, higher percentages of Indigenous students and lower percentages of LBOTE students. Broadly speaking, there are more striking differences between poorly performing schools and schools that have no dark red flags than between underperforming schools and schools with no dark red flags. This gives us some confidence that schools in our poorly performing

[15] The resulting standard error is essentially the square root of the weighted sum (weighted by the σ_j) of the variances and covariances among the individual β_j^i estimates within each sphere.

Table 1
Response Variable Means, Government Primary Schools, 2009

Policy sphere/variable	No dark red	0 < dark red < 20%	20%+ dark red
Policies to improve low-performing students			
Recommend grade retention	0.18	0.17	0.14
Provide additional tutoring in class	0.73	0.73	0.82
Provide additional tutoring outside class	0.06	0.02	0.04
Provide Saturday classes	0.00	0.00	0.00
Develop individual learning plans	0.97	0.99	0.94
Other policy not listed	0.42	0.40	0.38
Lengthening instructional time			
Average length of school day for middle grade (minutes)	371.9	372.7	368.3
Reduced class size for subject			
Regular students: Mathematics	0.40	0.35	0.42
Regular students: Reading	0.37	0.30	0.45
Regular students: Writing	0.31	0.27	0.39
Students with academic difficulties: Mathematics	0.54	0.52	0.49
Students with academic difficulties: Reading	0.65	0.59	0.54
Students with academic difficulties: Writing	0.45	0.48	0.50
Students with English as a second language: Mathematics	0.30	0.27	0.23
Students with English as a second language: Reading	0.38	0.29	0.26
Students with English as a second language: Writing	0.31	0.27	0.25
Narrowing of curriculum			
Minimum time required: Mathematics	0.77	0.84	0.68
Minimum time required: Writing	0.68	0.78	0.64
Minimum time required: Reading	0.75	0.75	0.68
No minimum required time: Science	0.55	0.63	0.53
No minimum required time: Art	0.44	0.51	0.54
No minimum required time: Social Studies	0.59	0.61	0.61
No minimum required time: Physical Education	0.28	0.24	0.27
Policies to improve low-performing teachers			
Increase supervision	0.80	0.81	0.78
Assign a teacher's aide	0.05	0.10	0.20
Assign a mentor or leading teacher	0.73	0.62	0.61
Provide additional professional development	0.89	0.94	0.91
Provide coaching from the principal	0.56	0.62	0.70
Other policy not listed	0.24	0.28	0.29
Teacher assigned time (hours/week)			
To collaboratively plan curriculum and assessment	1.13	1.08	0.90
To collaboratively review/monitor student performance	1.03	0.99	0.94
For class planning	2.40	2.58	2.04
Teacher control (1 = no influence, 4 = complete control)			
Establishing curriculum	2.89	2.90	2.83
Hiring new full-time teachers	1.98	1.87	1.53
Budget spending	2.59	2.60	2.55
Teacher evaluation	2.38	2.09	2.21
State control (1 = no influence, 4 = complete control)			
Establishing curriculum	3.24	3.37	3.26
Hiring new full-time teachers	2.99	3.27	3.57
Budget spending	2.43	2.59	2.62
Teacher evaluation	2.08	2.05	2.14
Principal control (1 = no influence, 4 = complete control)			
Establishing curriculum	2.92	3.01	2.83
Hiring new full-time teachers	2.86	2.71	2.43
Budget spending	3.25	3.22	3.07
Teacher evaluation	3.43	3.43	3.51

Table 1
(continued)

Policy sphere/variable	No dark red	0 < dark red < 20%	20%+ dark red
School climate			
Most parents closely monitor instructional program[†]	2.60	2.26	2.25
Most parents help children with homework[†]	2.79	2.52	2.05
Teachers recognised for improved student performance[†]	3.25	3.36	3.22
Require parents to sign children's homework (1 = yes)	0.34	0.28	0.29
Teachers have low expectations of students (reversed)[†]	3.73	3.49	3.32
Frequency principal interaction with parents: Phone[‡]	3.25	3.21	3.10
Frequency principal interaction with parents: In-person[‡]	3.57	3.55	3.45
Reduced class size for gifted students			
Mathematics	0.33	0.39	0.23
Reading	0.25	0.33	0.29
Writing	0.24	0.28	0.25
Teacher time spent outside school hours (minutes/week)			
On class preparation, grading, parent conferences, meetings	478.1	459.6	475.5
With students on activities: music, sport, tutoring, field trips	93.2	88.3	57.4
Teachers observed in the classroom			
Principal observed a teacher's lesson[§]	3.27	3.43	3.19
Specialist or leading teacher critiqued a teacher's lesson[§]	2.11	1.97	2.15
Another teacher observed a teacher's lesson[§]	2.75	2.74	2.48
Assistance in the classroom			
Teaching assistants, at least 1 hour/day	0.79	0.85	0.94
Parents/volunteers, at least 2 hours/week	0.90	0.89	0.85
Additional teacher, at least 1 day/week	0.60	0.65	0.71
Coach/lead teacher, at least 4 hours/week	0.28	0.27	0.33
Teaching assistants, at least 1 hour/day (all classrooms)	0.06	0.18	0.36
Parents/volunteers, at least 2 hours/week (all classrooms)	0.06	0.05	0.06
Additional teacher, at least 1 day/week (all classrooms)	0.10	0.11	0.15
Coach/lead teacher, at least 4 hours/week (all classrooms)	0.05	0.06	0.05
Homework time expected for tested subjects			
Average minutes/night: Mathematics	10.8	9.3	9.5
Average minutes/night: Reading	15.1	11.9	12.0
Average minutes/night: Writing	5.3	5.0	9.3

'Dark red' refers to the percentage of all NAPLAN test scores in a school being identified as substantially below other schools. If the score is more than 0.5 standard deviations below other schools, it is considered substantially below. These averages were constructed after weighting by the probability of responding to our survey by the state, location, sector and type of schools. [†]1 = strongly disagree, 4 = strongly agree. [‡]In an average week, 1 = none, 2 = 1–3, 3 = 4–6, 4 = 7–9, 5 = 10+. [§]In this academic year, 0 = never, 1 = once, 2 = 2–3 times, 3 = 4–5 times, 4 = 6+ times.

group, with 20% or more dark red flags, are schools that are struggling.

Table 2 tabulates average responses by all schools (not only government primary schools) to questions on our 2009 survey by school sector. Note that the questions in the spheres of *teacher incentives*, *assessment of teachers* and *teacher dismissal frequency* were only asked of non-government (Catholic and independent) schools. By comparison with government schools, independent schools are less likely to mandate minimum class time for almost all subjects, give teachers more time for planning and reviewing, are more likely to offer tutoring to low-performing students both in and outside of class, and are more likely to offer smaller classes to gifted students. By contrast, Catholic schools are more likely to mandate minimum time for almost all subjects compared with government schools. Non-government schools of both types are more likely than government schools to assign a mentor or leading teacher to assist low-performing teachers, to have a longer school day, and to have higher average homework time in tested subjects. The responses of both independent and Catholic schools show patterns indicating stronger

Table 2
Response Variable Means by School Sector, 2009

Policy sphere/variable	Government	Independent	Catholic
Policies to improve low-performing students			
Recommend grade retention	0.17	0.27	0.15
Provide additional tutoring in class	0.77	0.84	0.77
Provide additional tutoring outside class	0.12	0.31	0.16
Provide Saturday classes	0.00	0.03	0.01
Develop individual learning plans	0.97	0.95	0.98
Other policy not listed	0.38	0.34	0.35
Lengthening instructional time			
Average length of school day for middle grade (minutes)	372.7	390.3	383.4
Reduced class size for subject			
Regular students: Mathematics	0.35	0.38	0.38
Regular students: Reading	0.35	0.33	0.35
Regular students: Writing	0.31	0.33	0.32
Students with academic difficulties: Mathematics	0.51	0.66	0.55
Students with academic difficulties: Reading	0.57	0.62	0.61
Students with academic difficulties: Writing	0.49	0.59	0.54
Students with English as a second language: Mathematics	0.23	0.19	0.19
Students with English as a second language: Reading	0.29	0.22	0.26
Students with English as a second language: Writing	0.27	0.22	0.23
Narrowing of curriculum			
Minimum time required: Mathematics	0.79	0.52	0.89
Minimum time required: Writing	0.67	0.43	0.86
Minimum time required: Reading	0.72	0.51	0.86
No minimum required time: Science	0.49	0.61	0.34
No minimum required time: Art	0.51	0.57	0.33
No minimum required time: Social Studies	0.54	0.61	0.31
No minimum required time: Physical Education	0.25	0.37	0.16
Policies to improve low-performing teachers			
Increase supervision	0.78	0.80	0.83
Assign a teacher's aide	0.12	0.16	0.16
Assign a mentor or leading teacher	0.66	0.79	0.81
Provide additional professional development	0.89	0.89	0.93
Provide coaching from the principal	0.61	0.57	0.56
Other policy not listed	0.26	0.26	0.19
Teacher assigned time (hours/week)			
To collaboratively plan curriculum and assessment	1.05	1.39	1.19
To collaboratively review/monitor student performance	1.02	1.35	0.99
For class planning	2.87	4.20	3.04
Teacher control (1 = no influence, 4 = complete control)			
Establishing curriculum	2.80	2.88	2.78
Hiring new full-time teachers	1.77	2.03	1.98
Budget spending	2.54	2.29	2.26
Teacher evaluation	2.27	2.32	2.18
State control (1 = no influence, 4 = complete control)			
Establishing curriculum	3.22	2.69	2.85
Hiring new full-time teachers	3.15	1.12	1.44
Budget spending	2.49	1.40	1.65
Teacher evaluation	2.13	1.30	1.59
Principal control (1 = no influence, 4 = complete control)			
Establishing curriculum	2.87	3.06	2.89
Hiring new full-time teachers	2.66	3.45	3.48
Budget spending	3.20	3.23	3.32
Teacher evaluation	3.41	3.41	3.45

TABLE 2
(continued)

Policy sphere/variable	Government	Independent	Catholic
School climate			
Most parents closely monitor instructional program[†]	2.39	3.08	2.62
Most parents help children with homework[†]	2.43	2.70	2.90
Teachers recognised for improved student performance[†]	3.24	3.13	2.96
Require parents to sign children's homework (1 = yes)	0.29	0.54	0.56
Teachers have low expectations of students (reversed)[†]	3.51	3.80	3.68
Frequency principal interaction with parents: Phone[‡]	3.04	3.00	3.05
Frequency principal interaction with parents: In-person[‡]	3.27	2.91	3.23
Reduced class size for gifted students			
Mathematics	0.27	0.46	0.27
Reading	0.26	0.36	0.27
Writing	0.23	0.41	0.19
Teacher time spent outside school hours (minutes/week)			
On class preparation, grading, parent conferences, meetings	488.8	443.0	436.6
With students on activities: music, sport, tutoring, field trips	98.1	108.6	55.2
Teachers observed in the classroom			
Principal observed a teacher's lesson[§]	3.30	2.89	3.00
Specialist or leading teacher critiqued a teacher's lesson[§]	2.13	2.44	2.20
Another teacher observed a teacher's lesson[§]	2.77	2.94	2.72
Assistance in the classroom			
Teaching assistants, at least 1 hour/day	0.86	0.89	0.97
Parents/volunteers, at least 2 hours/week	0.74	0.78	0.71
Additional teacher, at least 1 day/week	0.63	0.62	0.66
Coach/lead teacher, at least 4 hours/week	0.29	0.32	0.34
Teaching assistants, at least 1 hour/day (all classrooms)	0.20	0.11	0.16
Parents/volunteers, at least 2 hours/week (all classrooms)	0.07	0.06	0.05
Additional teacher, at least 1 day/week (all classrooms)	0.13	0.13	0.10
Coach/lead teacher, at least 4 hours/week (all classrooms)	0.04	0.05	0.03
Homework time expected for tested subjects			
Average minutes/night: Mathematics	11.5	15.1	14.8
Average minutes/night: Reading	13.3	16.5	14.9
Average minutes/night: Writing	8.0	11.6	12.1
Teacher incentives			
Leadership position		0.55	0.42
Choice of class		0.05	0.04
Release time from teaching		0.22	0.17
Attendance at conferences and workshops		0.43	0.32
Other non-financial incentive not listed		0.19	0.13
Offer financial incentives of any form		0.35	0.11
Assessment of teachers (importance of)			
Direct observation by a school leader[¶]		3.53	3.47
Peer evaluation[¶]		3.05	3.02
Test scores of students (reversed)[¶]		2.77	2.67
External evaluation[¶]		1.94	2.00
Teacher dismissal frequency			
Dismissed or counselled a teacher to leave in the last 3 years		0.76	0.37

These averages were constructed after weighting by the probability of responding to our survey by the state, location, sector and type of schools. [†]1 = strongly disagree, 4 = strongly agree. [‡]In an average week, 1 = none, 2 = 1–3, 3 = 4–6, 4 = 7–9, 5 = 10+. [§]In this academic year, 0 = never, 1 = once, 2 = 2–3 times, 3 = 4–5 times, 4 = 6+ times. [¶]1 = not important / 4 = very important.

principal control, weaker state and teacher control, higher expectations of students, and higher levels of parental involvement than government schools.

Among non-government schools only, independent schools are also more likely to use various forms of teacher incentive than Catholic schools, with some form of financial incentive for high-performing teachers being provided in 12 per cent of Catholic schools and 35 per cent of independent schools. Independent schools are also more likely to have dismissed or counselled a teacher to leave in the past three years.

Tests of differences in spheres of policies and practices in 2009 by initial performance among government primary schools are provided in Table 3. These estimates were constructed by simply regressing the sphere indices on indicators for poorly performing and underperforming schools, as defined above. Consistent with the results in Table 1, we see that poorly performing schools – but not underperforming schools – stand out statistically in the spheres of *lengthening instructional time*, *low-performing teachers*, *teacher assigned time*, and all sub-spheres of *control (teacher*, *state* and *principal)*. Both poorly performing and underperforming schools stand out in terms of *school climate* and *assistance in the classroom*.

We constructed analogous tests for differences in sphere indices by initial performance using all schools, not just government primary schools (Coelli *et al.*, 2018). The differences observed in Table 3 were even more evident in those tests. We also constructed tests akin to Table 3 but splitting schools by quartiles of average normalised scores[16] rather than by per cent dark red flags. The same pattern of differences emerged in those tests. Finally, we constructed tests of differences in sphere indices across school sectors rather than by initial performance. As expected, these test results were consistent with the individual policy differences reported in Table 2.

TABLE 3
Differences by Performance, Government Primary Schools, Policy Sphere Indices, 2009

Policy sphere	0 < dark red < 20%	20%+ dark red
Policies to improve low-performing students	−0.011 (0.032)	−0.032 (0.027)
Lengthening instructional time	0.045 (0.108)	−0.196** (0.091)
Reduced class size for subject	−0.087 (0.170)	−0.051 (0.141)
Narrowing of curriculum	0.095 (0.085)	−0.059 (0.076)
Policies to improve low-performing teachers	0.048 (0.048)	0.103** (0.041)
Teacher assigned time (hours/week)	0.006 (0.131)	−0.171* (0.102)
Teacher control	−0.040 (0.134)	−0.207* (0.120)
State control	0.178 (0.116)	0.211** (0.103)
Principal control	−0.031 (0.129)	−0.225* (0.115)
School climate	−0.097* (0.050)	−0.178*** (0.043)
Reduced class size for gifted students	0.182 (0.194)	−0.035 (0.161)
Teacher time spent outside school hours	−0.044 (0.159)	−0.134 (0.139)
Teachers observed in the classroom	0.012 (0.072)	−0.080 (0.062)
Assistance in the classroom	0.090** (0.041)	0.202*** (0.035)
Homework time expected for tested subjects	−0.237* (0.132)	−0.089 (0.117)

Notes: Each row of estimates is based on a weighted regression of the policy sphere index measure on indicators for the percentage of NAPLAN test scores in the school in 2008 and 2009 flagged dark red – significantly below other schools – using schools with no dark red flags as the baseline (omitted) category. Robust standard errors on the coefficient estimates are provided in parentheses. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively.

(ii) Changes in Policies and Practices from 2009 to 2012

Moving on to an examination of how school policies and practices changed between 2009 and 2012, we begin by showing changes in individual item responses between the two surveys for government primary schools in Table 4. This

[16] These normalisations were constructed in the same manner as the measures reported at the bottom of Tables A1 and A2. Details of construction are provided in the notes for those tables.

Table 4
Change in Variable Means, Government Primary Schools, 2009–2012

Policy sphere/variable	No dark red	0 < dark red < 20%	20%+ dark red
Policies to improve low-performing students			
Recommend grade retention	−0.06	0.04	−0.03
Provide additional tutoring in class	−0.06	−0.04	−0.09
Provide additional tutoring outside class	0.00	0.01	0.03
Provide Saturday classes	0.00	0.00	0.00
Develop individual learning plans	0.02	−0.01	0.05
Other policy not listed	−0.02	−0.02	0.10
Lengthening instructional time			
Average length of school day for middle grade (minutes)	−0.04	−1.46	0.41
Reduced class size for subject			
Regular students: Mathematics	0.08	0.24	−0.09
Regular students: Reading	0.08	0.25	−0.06
Regular students: Writing	0.07	0.23	−0.05
Students with academic difficulties: Mathematics	0.02	0.06	0.07
Students with academic difficulties: Reading	−0.05	0.13	0.08
Students with academic difficulties: Writing	0.08	0.13	0.06
Students with English as a second language: Mathematics	0.06	−0.12	−0.19
Students with English as a second language: Reading	−0.01	−0.06	−0.12
Students with English as a second language: Writing	0.06	−0.03	−0.10
Narrowing of curriculum			
Minimum time required: Mathematics	0.08	0.00	−0.16
Minimum time required: Writing	0.09	−0.07	−0.31
Minimum time required: Reading	0.04	−0.07	−0.16
No minimum required time: Science	−0.05	−0.14	0.27
No minimum required time: Art	−0.04	0.00	0.30
No minimum required time: Social Studies	0.03	−0.07	0.25
No minimum required time: Physical Education	−0.03	0.06	0.31
Policies to improve low-performing teachers			
Increase supervision	0.07	0.02	0.05
Assign a teacher's aide	0.04	0.04	0.00
Assign a mentor or leading teacher	−0.01	0.04	0.14
Provide additional professional development	0.03	−0.05	−0.02
Provide coaching from the principal	0.08	0.10	0.03
Other policy not listed	0.06	0.10	−0.02
Teacher assigned time (hours/week)			
To collaboratively plan curriculum and assessment	−0.11	0.10	−0.15
To collaboratively review and monitor student performance	0.00	−0.11	−0.26
For class planning	0.35	0.11	0.49
Teacher control (1 = no influence, 4 = complete control)			
Establishing curriculum	−0.15	−0.15	−0.27
Hiring new full-time teachers	0.20	−0.13	0.19
Budget spending	0.04	−0.04	0.18
Teacher evaluation	−0.19	0.28	0.15
State control (1 = no influence, 4 = complete control)			
Establishing curriculum	−0.18	0.08	0.09
Hiring new full-time teachers	−0.44	−0.14	−0.24
Budget spending	−0.33	−0.13	−0.31
Teacher evaluation	−0.36	0.48	0.29
Principal control (1 = no influence, 4 = complete control)			
Establishing curriculum	−0.05	−0.22	0.09
Hiring new full-time teachers	0.08	0.21	0.43
Budget spending	0.10	−0.05	0.14

Table 4
(continued)

Policy sphere/variable	No dark red	0 < dark red < 20%	20%+ dark red
Teacher evaluation	0.08	0.10	−0.19
School climate			
Most parents closely monitor instructional program[†]	0.01	0.20	−0.16
Most parents help children with homework[†]	0.06	0.12	−0.03
Teachers recognised for improved student performance[†]	−0.30	−0.15	0.11
Require parents to sign children's homework (1 = yes)	0.00	0.07	0.03
Teachers have low expectations of students (reversed)[†]	−0.03	0.18	0.14
Frequency principal interaction with parents: Phone[‡]	0.00	−0.17	−0.06
Frequency principal interaction with parents: In-person[‡]	−0.06	−0.10	0.07
Reduced class size for gifted students			
Mathematics	0.17	−0.04	0.14
Reading	0.07	0.00	0.01
Writing	0.09	−0.03	0.00
Teacher time spent outside school hours (minutes/week)			
On class preparation, grading, parent conferences, meetings	90.6	52.1	121.9
With students on activities: music, sport, tutoring, field trips	64.6	30.4	−9.3
Teachers observed in the classroom			
Principal observed a teacher's lesson[§]	0.08	−0.26	−0.05
Specialist or leading teacher critiqued a teacher's lesson[§]	−0.02	−0.35	0.12
Another teacher observed a teacher's lesson[§]	0.05	0.04	0.08
Assistance in the classroom			
Teaching assistants, at least 1 hour/day	0.01	0.02	0.04
Parents/volunteers, at least 2 hours/week	0.03	−0.08	−0.05
Additional teacher, at least 1 day/week	−0.01	−0.06	−0.08
Coach/lead teacher, at least 4 hours/week	0.00	0.07	0.11
Teaching assistants, at least 1 hour/day (all classrooms)	0.03	0.01	0.08
Parents/volunteers, at least 2 hours/week (all classrooms)	−0.02	0.09	0.03
Additional teacher, at least 1 day/week (all classrooms)	0.04	0.02	0.04
Coach/lead teacher, at least 4 hours/week (all classrooms)	0.04	−0.02	0.11
Homework time expected for tested subjects			
Average minutes/night: Mathematics	−0.93	2.29	−2.04
Average minutes/night: Reading	−1.19	1.32	0.80
Average minutes/night: Writing	0.49	2.19	−8.81

Note: See Table 1.

table shows that poorly performing schools saw a larger reduction between 2009 and 2012 in the likelihood of minimum time being mandated for various subjects, including subjects directly tested by NAPLAN (such as reading and writing) and those not directly tested (such as science and physical education). School principals in both low-performance groups perceived an increase in state control of teacher evaluations compared to schools with no dark red flags, which saw a decrease in this measure in absolute terms. Principals in poorly performing schools also perceived a relative increase in their control over curriculum and hiring teachers, a relative increase in the recognition of teachers for student improvement, a relative decrease in parental monitoring of the instructional program, and a relative decrease in the average expected homework time for tested subjects. Relative to schools with no dark red flags, underperforming schools saw a relative reduction between 2009 and 2012 in the length of the school day and in the frequency of teachers being observed in the classroom, a relative increase in the use of smaller class sizes for regular students, and a relative increase in both teacher control of teacher evaluation and the average expected homework time for tested subjects.

In sum, the changes observed in our array of more than 60 policies and practices from 2009 to

2012 in struggling schools are mixed, sometimes in the 'wrong' direction in terms of what we might think intuitively would promote better student performance on NAPLAN tests, and frequently indistinguishable from the changes in these same policies and practices in schools receiving no dark red flags. When changes at struggling schools appear to be in the 'right' direction, they seem to relate more to teachers than to students.[17]

(iii) Did Worse Reports on My School Result in Improvements to Policies and Practices?

We now move on to a formal consideration of the central question of whether those schools that were revealed to have low relative performance when My School was first released responded differently. Here we employ the estimation strategy described above in Equations (1) and (2). We use indicators of percentage of dark red flags over the 2008 and 2009 years combined as our measures of relative performance. Results for government primary schools are presented in Table 5, using test score comparisons to both similar schools and all schools.

Table 5 shows little systematic evidence of stronger policy responses in the 'right' direction by struggling government primary schools than by other government primary schools. One potential exception is in the sphere of *narrowing the curriculum*, where we find that poorly performing government primary schools report a statistically significant increase relative to similar government primary schools in the highest-performing group. An examination of the results of our individual policy regressions shows that this effect – which is in the context of reductions at poorly performing schools in minimum time requirements for most subjects, both tested and non-tested, as shown in Table 4 – is driven by disproportionate drops in the likelihood of minimum time being allocated to non-tested subjects such as science.[18] Hence, this apparent devotion of relatively more time to tested subjects in poorly performing schools results not from more time being devoted to literacy and numeracy, but from even less time being devoted to non-tested subjects, in an environment in which the time allocated to every subject is declining relative to what is observed in higher-performing schools.[19]

There is some evidence that, relative to similar but better-performing schools, underperforming government primary schools – but not poorly performing schools – increased policies to improve *low-performing teachers*. When we delve within this sphere to individual policies and practices (see Coelli *et al.*, 2018), we find that this effect is driven by increasing the assignment of mentors and lead teachers, and by additional professional development. Underperforming government primary schools, however, saw relative reductions in *teacher assigned time* across the board, but mostly in terms of time to collaboratively review and monitor student performance. Relative to all other schools, poorly performing schools have responded by increasing *assistance in the classroom*. This was driven by relative increases in the use of teacher assistants, parents/volunteers and coaches/lead teachers.

While the inverse probability weighting we apply during estimation should compensate for any bias arising from heterogeneous responses by schools across states, locations, sectors and school types, it may not necessarily overcome potential sample selection bias. If responding to the survey is a direct function of the responses to the surveys (i.e. a function of the dependent variables in our regressions), then our regression estimates may still be biased. The direction of bias in this case is likely to be attenuation in regression coefficients (Goldberger, 1981). On the other hand, if responding is simply a function of the observable characteristics of schools, then our regression estimates will not be biased if we control for those observable characteristics, which in our case include the fully interacted set of indicators of state, location, sector and school type.

[17] We also tested whether there were aggregate changes over time in the sphere indices, separately by school sector (Coelli *et al.*, 2018). Changes were insignificantly different from zero in most cases, apart from increases in classroom assistance in all sectors, and increases in policies to improve low-performing teachers and teacher overtime hours in government schools.

[18] Results for individual policies and practices within a subset of spheres are provided in Coelli *et al.* (2018).

[19] As noted at the end of Appendix II, the principal leading the government primary school may have changed between surveys in over 30 per cent of cases. When we confine ourselves to estimating effects among those schools that we observe with the same principal in both survey years, the results are consistent with those reported in Table 5 (see Coelli *et al.*, 2018).

Table 5
Effect of Initial Performance on Policies and Practices, Government Primary Schools

	Similar schools		All schools	
Policy sphere	DR1	DR2	DR1	DR2
Policies to improve low-performing students	0.003	–0.042	0.014	0.056
	(0.048)	(0.072)	(0.057)	(0.065)
Lengthening instructional time	–0.007	0.044	–0.126*	–0.025
	(0.054)	(0.069)	(0.062)	(0.059)
Reduced class size for subject	–0.248	–0.267	–0.135	–0.096
	(0.166)	(0.251)	(0.200)	(0.202)
Narrowing of curriculum	0.136	0.273**	–0.050	0.173
	(0.100)	(0.127)	(0.108)	(0.120)
Policies to improve low-performing teachers	0.114**	0.000	0.057	0.109
	(0.053)	(0.086)	(0.064)	(0.070)
Teacher assigned time (hours/week)	–0.407**	–0.152	–0.100	–0.150
	(0.167)	(0.129)	(0.188)	(0.138)
Teacher control	–0.003	0.083	–0.172	–0.098
	(0.153)	(0.198)	(0.151)	(0.161)
State control	0.278*	0.008	0.289*	0.025
	(0.159)	(0.242)	(0.163)	(0.215)
Principal control	–0.114	–0.103	–0.284	–0.186
	(0.212)	(0.204)	(0.200)	(0.167)
School climate	–0.007	–0.061	–0.018	0.002
	(0.078)	(0.091)	(0.107)	(0.089)
Reduced class size for gifted students	–0.072	–0.355	0.149	0.045
	(0.210)	(0.246)	(0.272)	(0.252)
Teacher time spent outside school hours	–0.077	–0.203	–0.001	–0.106
	(0.144)	(0.290)	(0.166)	(0.194)
Teachers observed in the classroom	0.037	0.045	0.065	0.121
	(0.072)	(0.105)	(0.087)	(0.092)
Assistance in the classroom	0.091	–0.005	0.104*	0.146**
	(0.060)	(0.069)	(0.059)	(0.063)
Homework time expected for tested subjects	0.051	–0.165	0.045	–0.085
	(0.178)	(0.228)	(0.160)	(0.125)

Notes: Each pair of estimates in the table is drawn from a separate set of weighted seemingly unrelated regressions. See text for full details about the specifications employed. In all models, the dependent variable is the policy measure in 2012, and the same policy measure in 2009 is included as a covariate along with the relevant indicators of school performance (either with respect to all schools or with respect to similar schools). DR1 denotes 'dark red' proportions above 0 but below 0.2. DR2 denotes 'dark red' proportions of 0.2 and above. The full set of state by location fixed effects is included along with the following additional school-specific covariates: school size, Indigenous percentage, ICSEA score (all measured in 2009) and LBOTE percentage (measured in 2010). ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively.

One common approach to overcoming sample selection bias is to use Heckman's (1979) technique. This essentially regards sample selection bias as an omitted variable bias, and controls for it by including in the suite of independent variables a selection term, the inverse Mills ratio, based on estimates from an equation estimating the probability of sample selection. We take that approach and recover estimates (available in Coelli *et al.*, 2018) close to the ones reported in Table 5 and below.[20]

In a robustness exercise where we group schools into performance bands based on quartiles of initial scores rather than on dark red flags (Coelli *et al.*, 2018), we continue to find increases

[20] In the absence of reasonable instruments for selection, our selection term is identified by functional form alone.

among more poorly performing government primary schools in policies and practices related to teacher management, including observation and assistance in the classroom. As in our main results, we also find evidence of narrowing the curriculum, and no student-centred responses in the 'right' direction, although the relative likelihood of using smaller classes for gifted students falls.[21]

Having analysed government primary schools, Table 6 reports analogous results for all schools. In this broader sample, we again find some differences between regular and low-performing schools in the spheres of *narrowing the curriculum*, *teacher assigned time* and *assistance in the classroom*. We also analyse responses among the three spheres relevant only to non-government schools, as shown at the bottom of the table. Among these three spheres, there is a significant relative increase in *assessment of teachers* among poorly performing non-government schools.[22]

In Coelli *et al.* (2018) we report the results of estimating equations where we accommodate different performance-based responses for the three school sectors, by interacting the initial performance indicators with school sector. Results indicate that of the 7 (of 18 total) spheres in which poorly performing independent schools show statistically significant changes relative to better-performing independent schools, all changes except for an increase in teacher dismissals are in the 'wrong' direction (including fewer policies to improve poorly performing teachers, worse school climate, larger class sizes, reduced teacher time outside class hours, and decreased teacher assessment). Poorly performing Catholic schools show mixed responses, with some trends in the 'wrong' direction but also some reductions in class sizes and an increase in teacher assessments. Poorly performing government schools, by contrast, show increases relative to better-performing government schools in policies to improve poorly performing teachers, observation of teachers in the classroom, and assistance in the classroom. Despite the direction of change for poorly performing independent schools, underperforming independent schools show some relative changes in the 'right' direction – including teacher time spent outside school hours, and hours of homework assigned.

Our focus thus far has been on trying to understand whether schools that were revealed to be poorly or underperforming on My School responded in substantive ways. It may be of additional interest to understand how schools revealed to be high performing responded to My School. We perform an exercise parallel to Equations (1) and (2) to examine responses at the top end of the school performance distribution in Coelli *et al.* (2018). Results show that high-performing schools (with 20% or more dark green flags) were much less likely to have narrowed the curriculum to focus on tested subjects, more likely to have increased policies to improve low-performing teachers, less likely to feel that state control had increased, had an improved school climate, and were much less likely to have dismissed a teacher.

(iv) Subject Targeting

In an environment with binding budget constraints, principals may be unable to increase resources and emphasis on all NAPLAN-tested learning domains at the same time. However, if schools are performing worse in one tested area relative to another, constrained principals may reallocate resources towards the lower-performing subject area. To investigate whether this type of targeted response occurred, we combine the NAPLAN test score information and the responses of school principals with regard to subject-specific policies and practices into two separate areas: numeracy and literacy. NAPLAN covers one numeracy domain and four literacy domains: reading, grammar/punctuation, writing and spelling. In our surveys, we asked questions regarding one numeracy-related subject

[21] As noted in footnote 4, Tasmania and Western Australia had internet-based reporting of government school test score outcomes prior to My School. Our estimates of Table 5 are robust to the exclusion of schools from these two states. If My School is the driver of the small number of responses we observe in Table 5, we would expect that such responses should be less evident in Tasmania and Western Australia. However, due to small sample sizes (only 15 per cent of our 2012 responder schools are located in these two states), it was not possible to reliably test this hypothesis. We document our failed attempt in Coelli *et al.* (2018).

[22] Our robustness exercise of grouping schools into performance quartiles also yields similar results for all schools, with narrowing the curriculum and assistance in the classroom both on the relative rise among more poorly performing schools. In addition, the relative fall in school climate that is present but non-significant in our main results becomes significant when using quartiles to generate performance groups.

Table 6
Effect of Initial Performance on Policies and Practices, All Schools

	Similar schools		All schools	
Policy sphere	DR1	DR2	DR1	DR2
Policies to improve low-performing students	0.016	–0.018	0.025	0.040
	(0.029)	(0.046)	(0.037)	(0.040)
Lengthening instructional time	0.004	0.047	–0.255***	–0.014
	(0.069)	(0.091)	(0.077)	(0.077)
Reduced class size for subject	–0.246*	–0.008	–0.190	–0.021
	(0.131)	(0.258)	(0.179)	(0.178)
Narrowing of curriculum	0.075	0.222**	–0.060	0.112
	(0.091)	(0.113)	(0.098)	(0.102)
Policies to improve low-performing teachers	0.043	0.019	0.029	0.104*
	(0.042)	(0.074)	(0.049)	(0.059)
Teacher assigned time (hours/week)	–0.390***	–0.171*	–0.220	–0.096
	(0.105)	(0.089)	(0.154)	(0.112)
Teacher control	0.055	0.112	–0.132	–0.139
	(0.125)	(0.175)	(0.124)	(0.144)
State control	0.068	–0.053	0.219*	–0.089
	(0.114)	(0.195)	(0.132)	(0.183)
Principal control	0.081	–0.049	–0.219	–0.134
	(0.169)	(0.189)	(0.167)	(0.161)
School climate	–0.047	–0.108	–0.126	–0.080
	(0.066)	(0.081)	(0.082)	(0.079)
Reduced class size for gifted students	–0.084	–0.067	0.099	0.149
	(0.172)	(0.246)	(0.242)	(0.211)
Teacher time spent outside school hours	–0.099	–0.153	0.138	–0.047
	(0.140)	(0.279)	(0.164)	(0.188)
Teachers observed in the classroom	0.040	0.044	0.089	0.133*
	(0.057)	(0.088)	(0.068)	(0.074)
Assistance in the classroom	0.050	0.041	0.048	0.124**
	(0.048)	(0.059)	(0.046)	(0.053)
Homework time expected for tested subjects	–0.101	–0.307*	0.068	–0.121
	(0.126)	(0.166)	(0.127)	(0.123)
Teacher incentives	0.033	–0.223	–0.056	–0.216
	(0.083)	(0.175)	(0.072)	(0.136)
Assessment of teachers	–0.019	1.608***	0.152	0.436
	(0.187)	(0.160)	(0.215)	(0.428)
Teacher dismissal frequency	0.229	0.980	0.732*	0.365
	(0.316)	(0.742)	(0.400)	(0.564)

Notes: Each pair of estimates in the table is drawn from a separate set of weighted seemingly unrelated regressions. See text for full details about the specifications employed. In all models, the dependent variable is the policy measure in 2012, and the same policy measure in 2009 is included as a covariate along with the relevant indicators of school performance (either with respect to all schools or with respect to similar schools). DR1 denotes 'dark red' proportions above 0 but below 0.2. DR2 denotes 'dark red' proportions of 0.2 and above. The full set of state by location by sector by type fixed effects is included along with the following additional school-specific covariates: school size, Indigenous percentage, ICSEA score (all measured in 2009) and LBOTE percentage (measured in 2010). ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively.

(mathematics) and two literacy-related subjects (reading and writing).

Our measure of subject-specific performance for numeracy was constructed as the average of the numeracy × grade × cohort normalised test scores across all tested grades in a school and over the 2008 and 2009 cohorts (years). For literacy, we calculated the average of the domain × grade × cohort normalised test scores across the four literacy-related testing domains and over all tested grades in a school and over the 2008 and 2009 cohorts. Two sets of normalised measures were constructed, using raw scores for similar and for all schools when constructing the normalisations. We then defined our ‘relative performance’ measure (RP) as simply the numeracy score minus the literacy score, with a higher value on this measure indicating a stronger performance in numeracy relative to literacy.

We constructed our measure of stronger emphasis in policy spheres on numeracy relative to literacy (RE) based on principals’ responses on the following subject-specific questions, where all variables are indicators:

- Reduced class size for subject – regular students.
- Reduced class size for subject – students with learning difficulties.
- Reduced class size for subject – students from an English as a second language background.
- Minimum time required spent on subject each week.
- Typically, a minimum amount of time is spent on the subject each week.[23]

We combine responses to these subject-specific questions by normalising individual responses to have mean 0 and standard deviation 1 across all schools, and then constructing the simple average for each school of its normalised responses within each subject-specific category (mathematics for numeracy, and the combination of reading and writing for literacy). Our measure of relative emphasis in mathematics relative to literacy in terms of policies and practices is then simply the numeracy index minus the literacy index.

Our equation to estimate school-level policy responses in terms of relative emphasis on numeracy compared to literacy, based on school-level relative performance in numeracy compared to literacy, is

$$\mathrm{RE}_t = \alpha + \beta \cdot \mathrm{RP}_{t-1} + X_{t-1} \cdot \Psi + \varepsilon_t \qquad (3)$$

Here, RE_t is relative emphasis on numeracy compared to literacy subjects as revealed in the 2012 survey, RP_{t-1} is relative performance in numeracy versus literacy NAPLAN domains over 2008 and 2009, and X_{t-1} includes RE_{t-1} (relative emphasis in 2009) and a number of school-level indicators and characteristics measured in 2009: the relevant fully interacted set of state × location × sector × type indicators, ICSEA, LBOTE percentage, Indigenous percentage and enrolment count. The β coefficient is expected to be negative if schools respond to relatively poor performance in numeracy (literacy) by placing more emphasis on numeracy (literacy) relative to literacy (numeracy) in the policy arena.

Table 7

Effect of Initial Relative Performance on Subject-Specific Policies and Practices

	Comparator group	
	Similar schools	All schools
Government primary schools	−0.339** (0.152)	−0.283* (0.152)
All schools	−0.143 (0.202)	−0.083 (0.208)

Notes: Each estimate in the table is drawn from a separate ordinary least squares regression. See text for full details about the specifications employed. In all models, the dependent variable is the relative emphasis on numeracy versus literacy subjects in 2012. The coefficient reported is on the relative performance in numeracy versus literacy domains in the school averaged over 2008 and 2009. The other variables included in the regressions are the relative emphasis on numeracy versus literacy subjects in 2009, state by location by sector by type fixed effects (where relevant), school size, Indigenous percentage, ICSEA score (all measured in 2009) and LBOTE percentage (measured in 2010). White robust standard errors are provided in parentheses. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively.

The results from estimating Equation (3) are shown in Table 7. For government primary schools only, there is evidence of targeted policy responses to poor performance in one subject relative to the other in the expected direction. Such responses are not evident when we estimate the model for all schools together, which may indicate a more binding budget constraint for government schools than for independent and Catholic schools, and/or more responsiveness in poorly performing government schools than in poorly performing schools in other sectors to the demonstrated learning needs of their students.

(v) Principals’ Perceptions of the My School Website

Given the somewhat controversial nature of the My School website, particularly among teachers, we ended our 2012 survey by asking principals

[23] Yes/no answers to this question about specific subjects were collected only from random subsets of school principals using our additional survey modules.

whether they believed that the introduction of the My School website had a positive, negative or neutral effect on their school. Overall, 67 per cent of school principals responded that the My School website had a neutral effect on their school, 24 per cent said that it had a negative effect, and 8 per cent said it had a positive effect.[24] Across school sectors, government schools were least positive, while independent schools were most positive (see Coelli *et al.*, 2018 for details).

To investigate whether a school's reported performance on My School and its principal's perceptions of My School were related, we estimated ordered logit models of the three response values for the question about perception of My School (negative, neutral and positive) on schools' initial normalised scores on the NAPLAN tests. We included in a single model both normalised scores using the all-school comparisons and normalised scores using the similar-school comparisons, both calculated as averages over the 2008 and 2009 school years.

Results reveal that poor performance relative to similar schools was a key driver of principals' negative perceptions of the My School website (Coelli *et al.*, 2018). Principals of schools with low NAPLAN test scores relative to similar schools were more likely to report that the My School website had a negative effect on their school. Specifically, the principal of a school that was one standard deviation lower than average in terms of initial performance was 5 percentage points more likely to respond that My School had a negative effect. This finding is consistent with the conjecture that prior to My School, parents and other stakeholders may already have had a reasonable idea about a school's level of absolute performance, but that My School provided new information to parents about school performance relative to similar schools, potentially leading to uncomfortable conversations at the school level.

IV *Concluding Remarks*

Based on targeted surveys of school principals before and after the policy change, we generate the first evidence for Australia of the impact on schools' policies and practices from the one-shot increase in school accountability represented by the 2010 launch of Australia's My School website. In the study closest to ours, evaluating changes to Florida school accountability, Rouse *et al.* (2013) find that poorly performing schools 'are more likely to focus on low-performing students, lengthen the amount of time devoted to instruction, adopt different ways to organise the day and learning environment of the students and teachers, increase resources available to teachers, and decrease principal control'. By contrast, we find little systematic evidence of a pattern whereby schools that were revealed to have lower levels of performance systematically responded by changing their policies and practices relative to other schools in directions clearly aligned with improving student performance. While we do see some positive relative changes in policies and practices at struggling schools related to teacher support and incentives, we see almost no relative changes to student-focused policies and practices, and the direction of change in minimum class time and time assigned to teacher preparation is the opposite of what intuitively should support student learning. Despite observing few changes overall, we do observe the most positive trajectories of change in poorly performing government schools, and the least positive trajectories in poorly performing independent schools.

We also find mild evidence of policy targeting towards the learning domain (whether numeracy or literacy) on which a government primary school performed relatively worse, perhaps indicating the presence of binding resource constraints among such schools. We find that the typical principal perceived the My School website to have had a neutral effect on his or her school, with principals of lower-performing schools more likely than principals of other schools to report negative perceptions of test score reporting.

Is our evidence of weak accountability effects in response to My School explained by weak incentives for Australian principals, or by rigidities in the policy-setting environment? Freeman *et al.* (2014) draw on the 2013 edition of the OECD's Teaching and Learning International Survey (TALIS) data to report that 95 per cent of Australian school principals (compared to 89 per cent for OECD nations on average) stated that in the preceding 12 months, they had 'used student performance and student evaluation results (including national/international assessments) to develop the school's educational goals and programmes' (p. 45). This indicates a stance of above-average willingness on the part of

[24] Percentages do not sum to 100 due to rounding.

Australian principals to make changes to school practices in line with student performance data. This responsiveness could be driven by principals' career incentives, a desire to minimise complaints from parents and other stakeholders (discussed in Andrabi *et al.*, 2017, pp. 1546–7), and/or a simple desire to try to meet students' learning needs.

Evidence of Australian principals' ability to shape their schools can be drawn from that same TALIS survey only a few years earlier. Jensen (2010, p. 12) reports that Australia is the fourth lowest in the OECD in terms of the share of teachers who report that 'the most effective teachers [in their school] receive the greatest monetary or non-monetary rewards', and in terms of the proportion of teachers who believe they would receive some recognition if they were to improve the quality of their teaching, or (as a separate question) if they were to innovate in their teaching. Fewer than one in ten Australian teachers agreed with these each of these three statements separately. This indicates a possible breakdown in the chain from initial student performance and principals' intent to take responsive action, through to the implementation of responsive change, at least in policies and practices that relate directly to teacher performance.

Another possibility is that it takes time for school principals to adapt to a sudden change in public scrutiny. It may be the case that in a high-information environment such as the United States, school principals respond swiftly to changes in the perceived ranking of their school. Yet when a system such as Australia's moves from providing very little comparable school information to providing substantial information, it may take more than a few years for school principals to react. Notwithstanding the dramatic change in available test score information that occurred in 2010, developing a culture of responding to NAPLAN results may be something that occurs over decades.

Despite the caveats on our results – most importantly, our reliance on principals' choice of whether to respond to our surveys – the results of our surveys are directly relevant to education policy-makers. Our results indicate that poorly performing Australian schools have what appear to be worse policies and practices than other schools, and are falling behind in terms of time devoted to instruction and teacher preparation. Overall, our findings indicate that there is scope to improve struggling Australian schools via resourcing and policy decisions that better enable them to adapt their overall and student-centred policies and practices to improve outcomes for all students.

REFERENCES

ACARA (2014), 'Guide to Understanding 2013 Index of Community Socio-educational Advantage (ICSEA) Values', Fact Sheet. Australian Curriculum, Assessment and Reporting Authority, Sydney. [Cited 12 April 2018] Available from: http://docs.acara.edu.au/resources/Guide_to_understanding_2013_ICSEA_values.pdf

Andrabi, T., Das, J. and Khwaja, A.I. (2017), 'Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets', *American Economic Review*, **107**, 1535–63.

Booher-Jennings, J. (2005), 'Below the Bubble: "Educational Triage" and the Texas Accountability System', *American Educational Research Journal*, **42**, 231–68.

Carnoy, M. and Loeb, S. (2002), 'Does External Accountability Affect Student Outcomes? A Cross-State Analysis', *Educational Evaluation and Policy Analysis*, **24**, 305–31.

Chiang, H. (2009), 'How Accountability Pressure on Failing Schools affects Student Achievement', *Journal of Public Economics*, **93**, 1045–57.

Coelli, M. and Foster, G. (2016), 'Unintended Consequences of School Accountability Reforms: Evidence from Australia', Working Paper, University of Melbourne.

Coelli, M., Foster, G. and Leigh, A. (2018), 'Do School Principals Respond to Increased Public Scrutiny? New Survey Evidence from Australia', IZA Discussion Paper No. 11350, Institute of Labor Economics, Bonn.

Cullen, J. and Reback, R. (2006), 'Tinkering toward Accolades: School Gaming under a Performance Accountability System', in Gronberg, T. and Jansen, D. (eds), *Improving School Accountability* (Advances in Applied Microeconomics, Volume 14). Emerald Insight, Amsterdam; 1–34.

Dee, T.S. and Jacob, B. (2011), 'The Impact of No Child Left Behind on Student Achievement', *Journal of Policy Analysis and Management*, **30**, 418–46.

Deming, D.J., Cohodes, S., Jennings, J. and Jencks, C. (2016), 'School Accountability, Postsecondary Attainment, and Earnings', *Review of Economics and Statistics*, **98**, 848–62.

Figlio, D. (2006), 'Testing, Crime and Punishment', *Journal of Public Economics*, **90**, 837–51.

Figlio, D. and Getzler, L. (2006), 'Accountability, Ability and Disability: Gaming the System?' in Gronberg, T. and Jansen, D. (eds), *Improving School Accountability* (Advances in Applied Microeconomics, Volume 14). Emerald Insight, Elsevier JAI, Amsterdam; 35–49.

Figlio, D. and Ladd, H.F. (2015), 'School Accountability and Student Achievement', in Ladd, H.F. and Goertz, M.E. (eds), *Handbook of Research in Education Finance and Policy*. Routledge, New York, NY; 194–210.

Figlio, D. and Loeb, S. (2010), 'School Accountability', in Hanushek, E., Machin, S., and Woessmann, L. (eds), *Handbook of the Economics of Education*, Vol. 3. North-Holland, Amsterdam; 383–421. Available from: [Cited 12 April 2018] Available from: https://cepa.stanford.edu/sites/default/files/Accountability_Handbook.pdf.

Figlio, D. and Rouse, C. (2006), 'Do Accountability and Voucher Threats Improve Low-Performing Schools?', *Journal of Public Economics*, **90**, 239–55.

Figlio, D. and Winicki, J. (2005), 'Food for Thought: The Effects of School Accountability Plans on School Nutrition', *Journal of Public Economics*, **89**, 381–94.

Freeman, C., O'Malley, K. and Eveleigh, F. (2014) *Australian Teachers and the Learning Environment: An Analysis of Teacher Response to TALIS 2013: Final Report*. Australian Council for Educational Research, Melbourne.

Goldberger, A. (1981), 'Linear Regression after Selection', *Journal of Econometrics*, **15**, 357–66.

Haney, W. (2000), 'The Myth of the Texas Miracle in Education', *Education Policy Analysis Archives*, **8**, 41.

Hanushek, E.A. and Raymond, M.E. (2004), 'The Effect of School Accountability Systems on the Level and Distribution of Student Achievement', *Journal of the European Economic Association*, **2**, 406–15.

Heckman, J. (1979), 'Sample Selection Bias as a Specification Error', *Econometrica*, **47**, 153–61.

Helal, M. and Coelli, M. (2016), 'How Principals Affect Schools', Melbourne Institute Working Paper No. 18/16.

Hogan, W.J. and Lancaster, T. (2004), 'Instrumental Variables and Inverse Probability Weighting for Causal Inference from Longitudinal Observational Studies', *Statistical Methods in Medical Research*, **13**, 17–48.

Hussain, I. (2013), 'The School Inspector Calls', *Education Next*, **13**, 67–72.

Jacob, B.A. and Levitt, S. (2003), 'Rotten Apples: An Investigation of the Prevalence and Predictors of Teacher Cheating', *Quarterly Journal of Economics*, **118**, 843–77.

Jensen, P. (2010), 'What Teachers Want: Better Teacher Management', Grattan Institute Report 2010, No. 3. Grattan Institute, Carlton, Vic.

Kling, J. and Liebman, J. (2004), 'Experimental Analyses of Neighborhood Effects of Youth', Princeton University Industrial Relations Section Working Paper 483.

Lee, J. (2008), 'Is Test-Driven External Accountability Effective? Synthesizing the Evidence from Cross-State Causal-Comparative and Correlational Studies', *Review of Educational Research*, **78**, 608–44.

Lix, L.M. and Sajobi, T. (2010), 'Testing Multiple Outcomes in Repeated Measures Designs', *Psychological Methods*, **15**, 268–80.

Mizala, A. and Urquiola, M. (2013), 'School Markets: The Impact of Information Approximating Schools' Effectiveness', *Journal of Development Economics*, **103**, 313–35.

Neal, D. and Schanzenbach, D.W. (2010), 'Left Behind by Design: Proficiency Counts and Test-Based Accountability', *Review of Economics and Statistics*, **92**, 263–83.

OECD (2012), *Delivering School Transparency in Australia: National Reporting through My School, Strong Performers and Successful Reformers in Education*. OECD Publishing, Paris.

Patty, A. (2009), 'Schools Unite against Rankings', *Sydney Morning Herald*, 17 November.

Pugh, K. and Foster, G. (2014), 'Australia's National School Data and the "Big Data" Revolution in Education Economics', *Australian Economic Review*, **47**, 258–68.

Reback, R., Rockoff, J. and Schwartz, H.L. (2014), 'Under Pressure: Job Security, Resource Allocation, and Productivity in Schools under No Child Left Behind', *American Economic Journal: Economic Policy*, **6**, 207–41.

Rockoff, J. and Turner, L.J. (2010), 'Short-Run Impacts of Accountability on School Quality', *American Economic Journal: Economic Policy*, **2**, 119–47.

Rouse, C.E., Hannaway, J., Goldhaber, D. and Figlio, D. (2013), 'Feeling the Florida Heat? How Low-Performing Schools Respond to Voucher and Accountability Pressure', *American Economic Journal: Economic Policy*, **5**, 251–81.

West, M.R. and Peterson, P.E. (2006), 'The Efficacy of Choice Threats within School Accountability Systems: Results from Legislatively Induced Experiments', *Economic Journal*, **116**, C46–62.

Appendix I. School Characteristics

Table A1

School Characteristics by Respondent Status, 2009 Survey

				Controlled diff.	
Variable	Respondents	Non-respondents	*P*-value for difference	Amount	*P*-value
Number of students	378.4	383.0	0.612	−7.5	0.255
	(353.3)	(354.4)			
ICSEA score	1,004.7	1000.0	0.059	2.2	0.288
	(91.9)	(103.4)			
Indigenous (%)	6.89	8.47	0.0002	−0.72	0.024
	(14.54)	(17.99)			
LBOTE in 2010 (%)	16.50	17.33	0.178	0.03	0.956
	(23.02)	(23.87)			
Government (%)	70.3	71.3	0.414		
Catholic (%)	19.4	18.1	0.196		
Independent (%)	10.3	10.6	0.672		
Primary (%)	69.3	70.4	0.349		
Secondary (%)	16.5	15.2	0.180		
Combined (%)	14.2	14.4	0.871		
Metropolitan (%)	53.5	54.8	0.293		
Provincial (%)	39.2	38.2	0.451		
Remote (%)	4.6	3.7	0.073		
Very remote (%)	2.8	3.3	0.279		
Average normalised scores	−0.060	−0.070	0.442	0.008	0.465
	(0.448)	(0.532)			
Observations	1,872	7,279			

Notes: Special schools are excluded. All characteristics apart from language background other than English (LBOTE) are relevant to 2009, and most information is drawn from ACARA data. The average normalised scores were constructed by first normalising all school average test scores for each specific testing domain × grade × cohort grouping (e.g. reading results for students in grade 3 in 2009) by subtracting the overall Australian mean score for the same grouping and dividing by the overall Australian standard deviation. We then take the simple average of those normalised scores within a school for 2009. Tests of differences in characteristics were either *t*-tests of means (for quantitative variables) or *z*-score tests of proportions (for qualitative variables). Standard deviations are provided in parentheses. The controlled difference amount and *P*-value in the last two columns were constructed after controlling for state by sector by location by school type fixed effects using standard ordinary least squares regressions and heteroscedasticity-robust standard errors. Data are not available for all schools, as ACARA does not provide information for schools with extremely small student numbers. In such cases, we add in data on sector, type and location for all small schools from information provided by DEEWR or from our own collection efforts; the number of observations for which data on other variables are available differs by variable but ranges from 87% (for average normalised scores) to 97% (number of students) of the count of total observation shown in the final row of the table.

Table A2
School Characteristics by Respondent Status, 2012 survey (among 2009 Respondents)

				Controlled difference	
Variable	Respondents	Non-respondents	*P*-value for difference	Amount	*P*-value
Number of students	408.8 (369.4)	361.6 (348.1)	0.006	11.1	0.392
ICSEA score	1,008.5 (89.1)	997.2 (94.7)	0.010	0.21	0.951
Indigenous (%)	6.95 (14.33)	8.40 (15.25)	0.050	−0.10	0.853
LBOTE (%)	17.43 (23.84)	15.02 (21.88)	0.026	0.12	0.900
Government (%)	66.8	74.5	0.0003		
Catholic (%)	22.4	15.8	0.0004		
Independent (%)	10.8	9.7	0.4500		
Primary (%)	66.8	72.6	0.008		
Secondary (%)	18.4	13.8	0.010		
Combined (%)	14.9	13.6	0.436		
Metropolitan (%)	56.1	49.4	0.004		
Provincial (%)	36.9	42.6	0.014		
Remote (%)	4.4	4.9	0.652		
Very remote (%)	2.5	3.2	0.396		
Average normalised scores	−0.067 (0.47)	−0.110 (0.48)	0.071	−0.014	0.478
Observations	1,062	780			

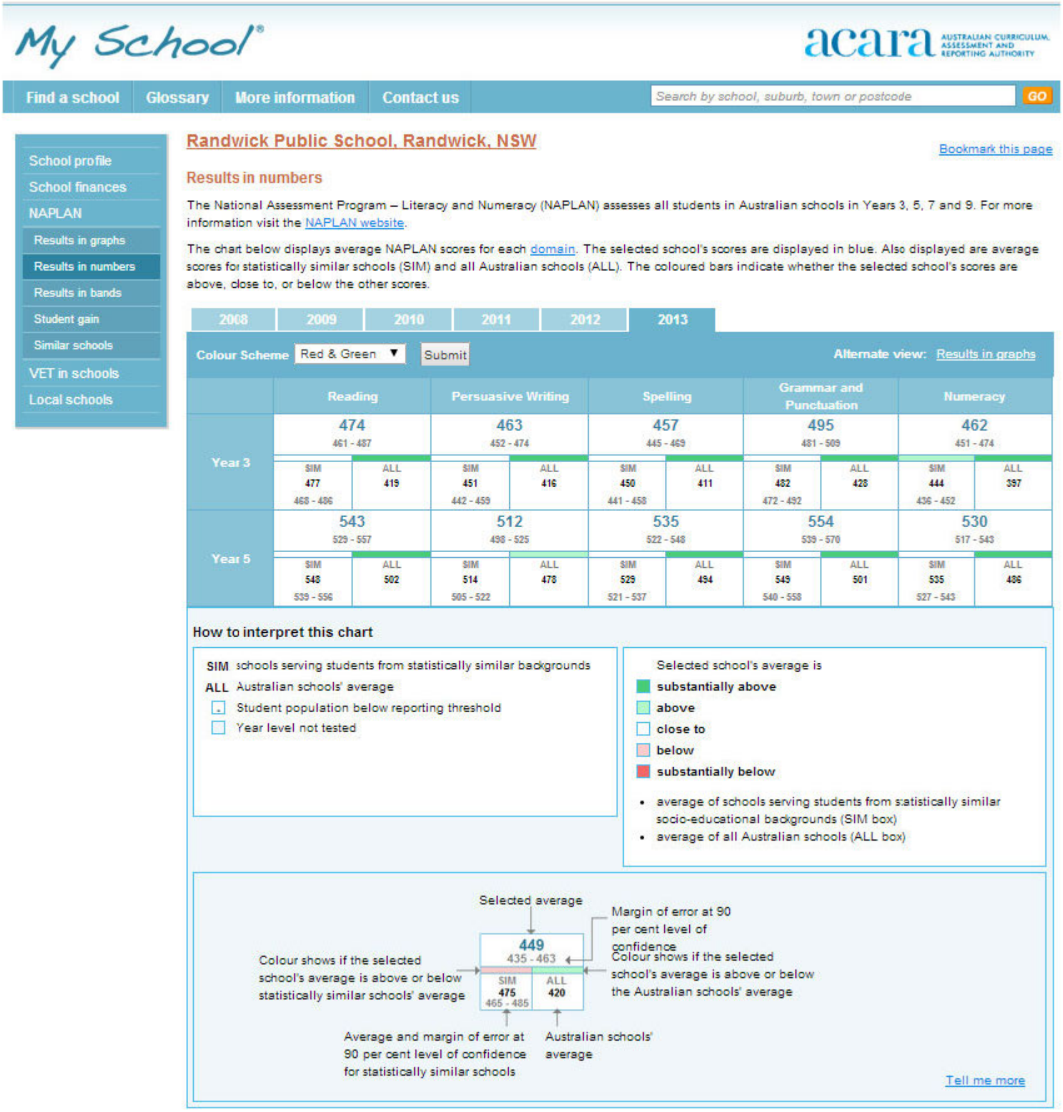
Notes: Special schools were excluded. All characteristics are 2012 measures from ACARA. LBOTE = language background other than English. The average normalised scores were constructed by first normalising all school average test scores for each specific testing domain × grade × cohort grouping (e.g. reading results for students in grade 3 in 2012) by subtracting the overall Australian mean score for the same grouping and dividing by the overall Australian standard deviation. We then take the simple average of those normalised scores within a school for 2012. Tests of differences in characteristics were either *t*-tests of means (for quantitative variables) or *z*-score tests of proportions (for qualitative variables). Standard deviations are provided in parentheses. The controlled difference amount and *P*-value in the last two columns were constructed after controlling for state by sector by location by school type fixed effects using standard ordinary least squares regressions and heteroscedasticity-robust standard errors. Data are not available for all schools, as ACARA does not provide information for schools with extremely small student numbers. In such cases, we add in data on sector, type and location for all small schools from information provided by DEEWR or from our own collection efforts; the number of observations for which data on other variables are available differs by variable but ranges from 89% (for average normalised scores) to 98% (LBOTE) of the count of total observation shown in the final row of the table.

TABLE A3
Government Primary School Characteristics by Per Cent Dark Red

Variable	No dark red	0 < dark red < 20%	20%+ dark red
Number of students	360.1	279.8***	238.1***
	(215.8)	(203.8)	(190.8)
ICSEA score	1,051.2	997.8***	945.6***
	(67.4)	(47.5)	(79.6)
Indigenous (%)	2.43	4.70***	12.70***
	(3.17)	(5.46)	(14.77)
LBOTE in 2010 (%)	22.6	16.3***	11.5***
	(23.8)	(24.1)	(18.7)
Observations	377	176	291

Notes: Per cent dark red based on comparisons with all other schools. All characteristics apart from language background other than English (LBOTE) are relevant to 2009, and all information is drawn from ACARA data. ***, ** and * denote statistical significance at the 1%, 5% and 10% levels, respectively, of tests of the differences between the group of schools in the given column and schools that have no dark red indicators, based on *t*-tests (all estimated using weights).

Appendix II. Screenshots of My School NAPLAN Results Page for Two Schools [Colour figure can be viewed at wileyonlinelibrary.com]



Randwick Public School, Randwick, NSW

Results in numbers

The National Assessment Program – Literacy and Numeracy (NAPLAN) assesses all students in Australian schools in Years 3, 5, 7 and 9. For more information visit the NAPLAN website.

The chart below displays average NAPLAN scores for each domain. The selected school's scores are displayed in blue. Also displayed are average scores for statistically similar schools (SIM) and all Australian schools (ALL). The coloured bars indicate whether the selected school's scores are above, close to, or below the other scores.

2008
2009
2010
2011
2012
2013
Colour Scheme Red & Green ▼
Submit
Alternate view: Results in graphs

	Reading		Persuasive Writing		Spelling		Grammar and Punctuation		Numeracy	
Year 3	474 461 - 487		463 452 - 474		457 445 - 469		495 481 - 509		462 451 - 474	
	SIM 477 468 - 486	ALL 419	SIM 451 442 - 459	ALL 416	SIM 450 441 - 458	ALL 411	SIM 482 472 - 492	ALL 428	SIM 444 436 - 452	ALL 397
Year 5	543 529 - 557		512 498 - 525		535 522 - 548		554 539 - 570		530 517 - 543	
	SIM 548 539 - 556	ALL 502	SIM 514 505 - 522	ALL 478	SIM 529 521 - 537	ALL 494	SIM 549 540 - 558	ALL 501	SIM 535 527 - 543	ALL 486

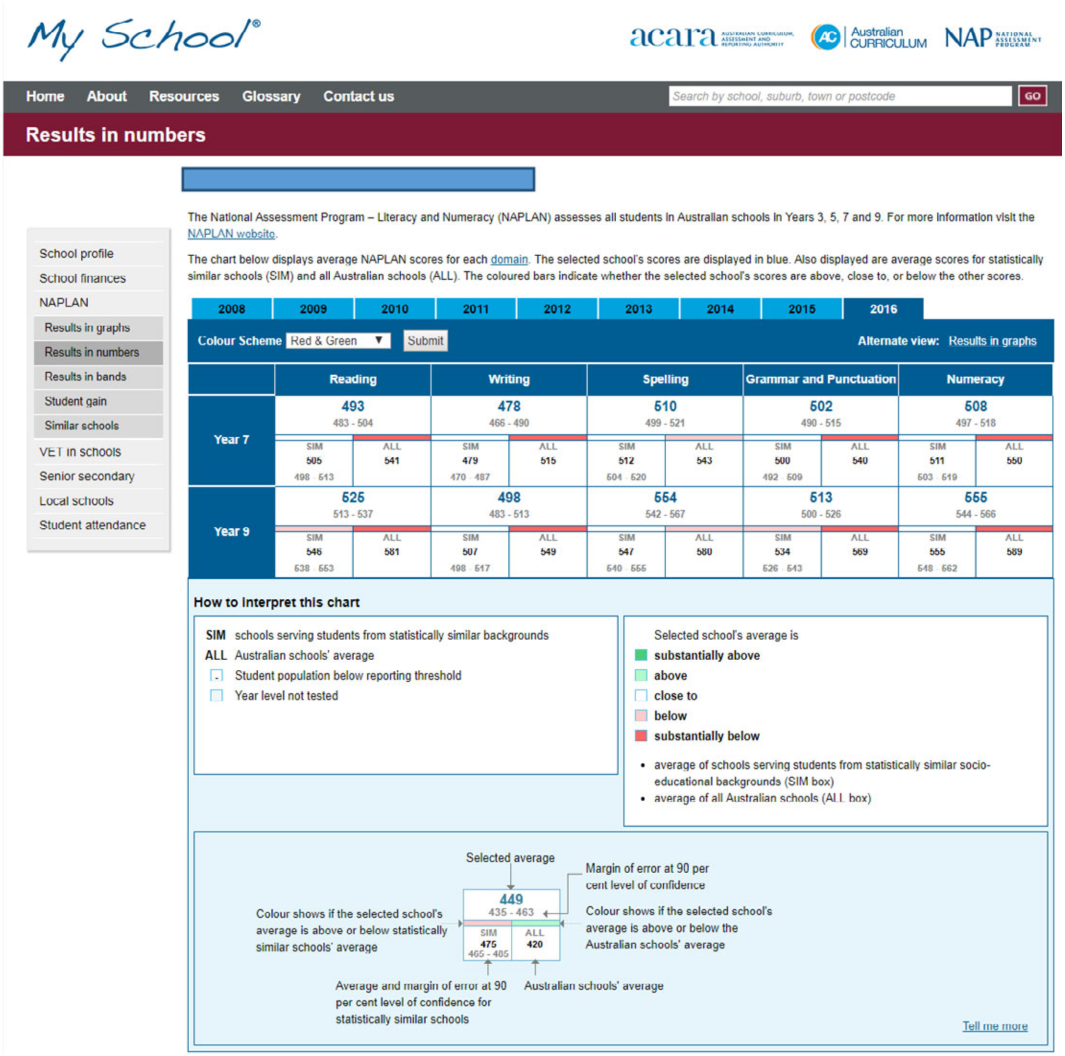
How to interpret this chart

SIM schools serving students from statistically similar backgrounds
ALL Australian schools' average
▫ Student population below reporting threshold
□ Year level not tested

Selected school's average is
🟩 substantially above
🟩 above
□ close to
🟥 below
🟥 substantially below

- average of schools serving students from statistically similar socio-educational backgrounds (SIM box)
- average of all Australian schools (ALL box)

Selected average
449
435 - 463
SIM
475
465 - 485
ALL
420
Margin of error at 90 per cent level of confidence
Colour shows if the selected school's average is above or below the Australian schools' average
Colour shows if the selected school's average is above or below statistically similar schools' average
Average and margin of error at 90 per cent level of confidence for statistically similar schools
Australian schools' average
Tell me more



My School®

Home | About | Resources | Glossary | Contact us

Search by school, suburb, town or postcode GO

Results in numbers

- School profile
- School finances
- NAPLAN
  - Results in graphs
  - Results in numbers
  - Results in bands
  - Student gain
  - Similar schools
- VET in schools
- Senior secondary
- Local schools
- Student attendance

The National Assessment Program – Literacy and Numeracy (NAPLAN) assesses all students in Australian schools in Years 3, 5, 7 and 9. For more information visit the NAPLAN website.

The chart below displays average NAPLAN scores for each domain. The selected school's scores are displayed in blue. Also displayed are average scores for statistically similar schools (SIM) and all Australian schools (ALL). The coloured bars indicate whether the selected school's scores are above, close to, or below the other scores.

2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016

Colour Scheme: Red & Green ▼ Submit

Alternate view: Results in graphs

	Reading		Writing		Spelling		Grammar and Punctuation		Numeracy	
Year 7	493 483 - 504		478 466 - 490		510 499 - 521		502 490 - 515		508 497 - 518	
	SIM 505 498 - 513	ALL 541	SIM 479 470 - 487	ALL 515	SIM 512 504 - 520	ALL 543	SIM 500 492 - 509	ALL 540	SIM 511 503 - 519	ALL 550
Year 9	525 513 - 537		498 483 - 513		554 542 - 567		513 500 - 526		555 544 - 566	
	SIM 546 538 - 553	ALL 581	SIM 507 498 - 517	ALL 549	SIM 547 540 - 555	ALL 580	SIM 534 526 - 543	ALL 569	SIM 555 548 - 562	ALL 589

How to interpret this chart

SIM schools serving students from statistically similar backgrounds

ALL Australian schools' average

□ Student population below reporting threshold

□ Year level not tested

Selected school's average is

🟩 **substantially above**

▨ **above**

□ **close to**

▨ **below**

🟥 **substantially below**

- average of schools serving students from statistically similar socio-educational backgrounds (SIM box)
- average of all Australian schools (ALL box)

Selected average
449
435 - 463
Margin of error at 90 per cent level of confidence
Colour shows if the selected school's average is above or below statistically similar schools' average
Colour shows if the selected school's average is above or below the Australian schools' average
SIM
475
465 - 485
ALL
420
Average and margin of error at 90 per cent level of confidence for statistically similar schools
Australian schools' average

Tell me more

Appendix III. Survey Implementation

(i) 2009 Survey

The initial survey was undertaken in the second half of 2009, with responses collected from 11 October 2009 until 29 January 2010 (only 1.4 per cent of responses were collected in January 2010, and all prior to the start of the 2010 school year). The entire population of Australian schools was included in the initial survey frame. The list of contact details for schools was provided by the Commonwealth Department of Education, Employment and Workplace Relations (DEEWR). This list included both government/public schools and private schools (Catholic and independent), and covered all school levels (primary, secondary and combined) including special schools (schools for children with learning disabilities).

School principals were initially sent a letter inviting them to complete our survey. The letter included a link to a website where the survey could be completed online. The letter also included a six-digit school-specific code provided by us that the school principal was required to enter in order to complete the survey. This school-specific code allowed us to track completion closely. Follow-up emails were sent to schools several days after the mailing of the initial letter. Schools that had not responded to the survey after the first contact were recontacted up to two more times spaced approximately one month apart via letters, telephone calls and follow-up emails to improve response rates.

There were five different versions of the survey sent to schools via random allocation. All five versions had a standard set of questions (a core module), with four of the five versions also having a small number of additional questions (additional modules). We chose to use several versions of the survey in order to reduce the response burden of individual school principals. Our aim was to keep survey completion time below 25 minutes.

Certain survey questions were only asked of private schools (Catholic and independent), as they were most relevant for those schools (tuition fees charged, incentives provided to teachers, etc.).

The response rate for the initial survey was approximately 21 per cent. In total, 1,959 schools completed the 2009 survey. In the vast majority of cases (96%), the school principal completed the survey. Another member of the school leadership team (deputy principal, registrant) answered on the school's behalf in the remaining 4% of cases.

(ii) 2012 Survey

The follow-up survey in 2012 was undertaken in the second half of 2012, with responses collected from 22 July 2012 until 20 December 2012. All 1,959 schools that responded to the initial survey in 2009 were approached to complete this second survey, but 30 of those initial responders had closed or merged with other schools in the intervening period. Thus 1,929 schools still operating as separate entities were potentially able to complete the 2012 survey. Schools were sent the same version of the survey that they completed in 2009, allowing the tracking of responses over time. The 2012 survey included a small number of new questions (not in the 2009 survey) specifically about the My School website. This website was not brought online until 2010.

School principals were again sent letters inviting them to complete our survey online, and again follow-up emails were sent several days later. Initial non-respondents were sent reminder letters and emails up to three more times (early in September, October and November). All schools that completed the 2012 survey were entered into a prize draw (if they chose to do so) for an education support package of the choosing of the school up to a value of $2,000. This prize draw was offered as an extra inducement to improve response rates for the follow-up survey.

As an extra measure to improve response rates, non-respondent schools were contacted by phone by the research team during the second half of November and early December in 2012. In many of the cases where a survey was completed in response to these phone calls, school principals answered the questionnaire directly over the phone rather than via the internet. Phone completions comprised just less than 5 per cent of all 2012 survey completions.

The response rate for the 2012 follow-up survey was approximately 58 per cent (after removal of schools that had closed or merged prior to 2012). In total, 1,122 schools completed the 2012 survey, at least partially. For this 2012 survey, 93% of responses were completed by the school principal.

Given that the school principal was not always the responder to our two surveys, it is not elementary to determine from our data whether a school experienced a change in principal between survey years. The principal was the responder to both surveys in 89% of schools. In 2% of these cases, the responding principal(s) did not provide their name(s) in at least one of the two surveys. Of schools with named, responding principals in both survey years, around 71% (i.e., about 62% of our entire school sample) had the same principal responding in both surveys. This 'same principal' proportion varied from a low of 68 per cent in government schools and 76 per cent in Catholic schools, to 84 per cent in independent schools. For comparison, Helal and Coelli (2016) found that the rate of year-to-year principal changes among Victorian government primary schools was around 15%, a little higher than the change over three years we observe of 32%.