Economics of Education Review 99 (2024) 102523

Contents lists available at ScienceDirect

Economics of Education Review

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

Unintended consequences of school accountability reforms: Public versus private schools

Michael Coelli [a,*], Gigi Foster [b]

[a] University of Melbourne, Department of Economics, Level 4, FBE Building, 111 Barry Street, Carlton, VIC, 3053, Australia
[b] University of New South Wales, Department of Economics, UNSW Australia School of Business Building, Kensington Campus, High Street, Kensington, NSW, 2052, Australia

ARTICLE INFO

JEL codes:
I21
I28

Keywords:
School accountability
Assessment
Literacy
Numeracy

ABSTRACT

In this paper, we show that the public provision of information on Australian schools' average national test score outcomes via the *My School* website, launched in 2010, resulted in poorly performing schools testing fewer students in subsequent years. This increased non-participation in testing was driven primarily by formal parental withdrawal, and poorly performing students were much more likely to be withdrawn from testing. This phenomenon is consistent with schools attempting to 'game' the system to improve published test scores. We also provide weaker evidence that withdrawal responded more strongly to initial poor performance in independent private schools than in government schools.

1. Introduction

The policy of publicly disseminating information on schools' test performance has several objectives. One is to increase the overall public accountability of schools, and a second is to equip parents with school quality information relevant to their enrolment decisions. Through both pathways, this type of policy seeks to increase the fraction of students attending higher-quality schools and thus achieving better outcomes.

The public spotlight, however, can motivate schools to attempt to raise their measured average test scores by means other than improving the quality of teaching. Schools may concentrate on high-stakes rather than low-stakes subjects; disproportionately teach skills that are valuable for the tests; and/or intervene disproportionately with students more likely to make a marginal difference to the school's performance (Figlio, Loeb, Hanushek, Machin & Woessmann, 2011). Schools may also attempt to boost test scores artificially (Figlio & Winicki, 2005), use selective discipline (Figlio, 2006), or even engage in outright cheating (Jacob & Levitt, 2003). They may also selectively exclude students from testing (Cilliers, Mbiti & Zeitlin, 2021; Cullen & Reback, 2006; Figlio & Getzler, 2006).

In this paper, we investigate how the percentage of students participating in national tests in Australian schools responded to the sudden increase in school accountability brought about by the launch of the *My School* website in January 2010. This website displays average test scores for each school, in each assessed cohort (grades 3, 5, 7 and 9), in five distinct learning domains, from 2008 onwards. As an advance on previous work, we can examine and compare the responses of both public and private schools, because *My School* affected all sectors in a manner free of the jurisdictional complexities that exclude private schools in many other countries from accountability mechanisms.[1]

Private schools provide choice for parents and competition to public schools. *Ex ante* we might expect a stronger response to public accountability by fee-charging private schools, given that on average they face better-off parents likely to have more schooling options and to be more focussed on student outcomes. Such schools must compete for and justify their tuition costs and may face heightened pressure from parents if school performance does not match expectations.

We find that schools in both the public and private sectors appear to respond strategically to the increase in accountability. The worse a school's initial measured performance as advertised on *My School*, the more likely that a lower proportion of students from that school participated in subsequent testing years. Evidence further suggests that

* Corresponding author.
E-mail address: mcoelli@unimelb.edu.au (M. Coelli).

[1] A handful of papers have also compared responses of both public and private schools to public accountability interventions in other jurisdictions: Andrabi et al (2017) in Pakistan, Carmargo et al. (2018) and Firpo et al. (2014) in Brazil, and Mizala and Urquiola (2013) in Chile. These studies focus on test score, enrolment and fee responses rather than on manipulation of the testing pool as we do.

https://doi.org/10.1016/j.econedurev.2024.102523
Received 28 September 2022; Received in revised form 21 February 2024; Accepted 27 February 2024
Available online 9 March 2024

students who do not participate in testing are more likely to be low performers. These responses threaten the efficacy of the public accountability mechanism in Australia and undermine the information provided to parents making enrolment decisions.

The outline of the paper is as follows. We provide details of the Australian education system and the introduction of the *My School* website in Section 2. We describe our data and provide summary statistics in Section 3. We describe our estimation strategy in Section 4. Our main estimates are reported and discussed in Section 5, and some concluding observations are offered in Section 6.

2. Background: the Australian education sector and national testing system

In Australia, public (government) schools are funded and run separately by Australia's six states and two territories, and do not charge fees. Private schools include Catholic and independent schools, with many of the latter also associated with religious organisations. Private schools charge fees and receive government funding that was, during the period analysed in this paper, on average equivalent to 70 % of the per-student government funding received by public schools. Standard public schools are required to accept students who reside in their catchment areas. Private schools and a small number of selective public schools, of which most are academically selective, have more control over the students they accept.

Most Australian schools offer either primary or secondary education. Primary (elementary) schools generally offer education from the year prior to grade 1[2] up to grade 6 or 7, depending on the state. Secondary (high) schools generally offer education from grade 7 or 8 to grade 12, the final year of formal schooling in Australia. Some schools – denoted "combined" schools – offer education over both primary and secondary levels.

To receive government funding, Australian private schools must be organised as not-for-profit enterprises. Our conjecture that private schools may respond more than public schools to public accountability is thus not based on a profit motive. The objective function of school leaders is likely multi-faceted, with one component being to meet or exceed the performance expectations of parents. We conjecture that parents are likely to set higher performance expectations for private schools – particularly those charging high fees – than for public schools.

In 2008 – two years prior to the launch of the *My School* website that publicised schools' average test results – Australian students began taking national tests in the five domains of reading, writing, spelling, grammar/punctuation and numeracy in grades 3, 5, 7 and 9, via the National Assessment Program – Literacy and Numeracy (NAPLAN). Testing is conducted by state school authorities but administered centrally by the Australian Curriculum and Reporting Authority (ACARA). Prior to NAPLAN, individual states conducted their own literacy and numeracy tests. NAPLAN testing occurred in May of each year during the period we examine, about one-third of the way through the academic year. Appendix A provides a sketch of the yearly testing and reporting timelines.

It should not be underestimated how contentious the introduction of national testing and reporting was in Australia. In 2009, there was an attempt to ban NAPLAN outright within New South Wales which brought down one of that state education department's Directors General. This occurred in the very state that had the longest and richest existing culture of testing students and using those test results to inform policy.

Ministers of Education in each state and territory responded to the *My School* project by putting together a special advisory group that made recommendations about how to amend or enhance the website. This advisory group included the heads of principals' councils, university academics, statisticians, and others, and worked with ACARA until August 2010. Many recommendations for changes were made, not all of which were implemented.

Schools were informed in broad terms about *My School* and the type of information it would reveal prior to the website launch. However, as made clear by the continued consultative process described above, what would be reported on *My School* and how exactly it would be displayed could reasonably be considered as in flux until mid-2009, and the initial version of the website itself was not made live until January 2010. Hence, actionable information is unlikely to have reached schools until after NAPLAN testing in May 2009. Any responses by schools to the change in accountability represented by *My School* are thus unlikely to be evident prior to the 2010 testing year.

The *My School* website went live on 28 January 2010, near the first day of school in the 2010 school year (school start dates differ slightly by state/territory). Schools' average test score information for both 2008 and 2009 was provided on the website at this time. Average test score information has subsequently been added annually to *My School* each March,[3] several months after individual schools and parents are sent their own test score information (around September of each testing year).

Each school's scores are compared on *My School* to two sets of average scores within testing domain × grade × cohort: (1) averages over all schools in Australia offering the same grade, and (2) averages for a group of up to 60 "similar" schools. These "similar" schools are identified by ACARA using their Index of Community Socio-Educational Advantage (ICSEA). The ICSEA score of each school is now based on the education and occupation levels of the parents of enrolled students, the remoteness of the school, and the percentage of enrolled students from an Indigenous background.[4] Comparisons with similar schools essentially hold constant these background characteristics in evaluating schools' test results.

Appendix B provides two screenshots from *My School*: 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 appropriate color coding, described below, for each of the five domains of testing (numeracy, reading, grammar/punctuation, writing and spelling) and for each of the grades (denoted "year") tested at the school. Test score comparisons are shown in the form of numbers and colours against both the "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 color of each indicator is determined by the distance from the school's score in that domain × grade × cohort to the mean score of the relevant comparison group. 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. Colourlessness denotes a score that is approximately the same as the comparison-group mean. The thresholds that define "substantially" and "somewhat" higher or lower than the comparison-group mean score are

[2] This grade has different names across states: kindergarten, preparation, or foundation. Students generally start school in this grade at age five or six.

[3] For example, schools' average scores for tests taken in May 2010 were added in March 2011.

[4] These three components are combined to form the ICSEA measure in a manner that best predicts student test scores. Prior to 2010, ICSEA was based on the socio-economic status of the postcode where students resided, as information on parental education and occupation has been collected since 2010 only. For more details: http://www.acara.edu.au/verve/_resources/Guide_to_understanding_2013_ICSEA_values.pdf.

that mean score plus or minus 0.5 and 0.2, respectively, of the standard deviation of scores in that domain × grade × cohort across all students in Australia.[5]

According to the National Protocols for Test Administration,[6] there are three reasons why a student may not participate in NAPLAN testing: (1) **exemption** due to diagnosed learning difficulties, or residence in Australia for less than one year and being from a language background other than English; (2) **absence** from school on testing day; or (3) being **withdrawn** from testing by formal parental application. Unlike in some jurisdictions (e.g., Florida, see Florida Department of Education, 2014), Australia sets no minimum percentage of students in each school and grade who must participate in testing.

Parents can formally apply to the school principal for the withdrawal of their child from NAPLAN testing "in the manner specified by the local testing authority" (usually the state's department of education). The NAPLAN testing protocols indicate that withdrawals are intended to address issues "such as religious beliefs and philosophical objections to testing," but providing a considered explanation for withdrawal does not appear necessary. Recent examples of forms used by local testing authorities for withdrawal of students from NAPLAN testing are provided in Appendix C.

The process of applying for withdrawal was not actively promoted by governments or the testing authority. We believe that knowledge of the process was provided to parents directly by schools. As noted above, parents apply for withdrawal directly to the school, not to the testing authority or government.

Parents have reported pressure from schools to withdraw children from testing in a survey related to a review of NAPLAN conducted by one state education authority.[7] The popular press has also reported on claims by parents that schools have instructed children not to sit the NAPLAN tests "…in order to boost their chances of obtaining higher overall scores…".[8] These reports date back as far as 2010, the first year of testing after the launch of *My School*. The volume of such reports led the NSW Minister for Education to warn that principals or teachers found encouraging children not to sit the tests may face disciplinary action. Reports of parents coordinating the withdrawal of their children from testing so their school would not be included on the *My School* website have also been made.[9]

In the Australian setting prior to *My School*, parents likely already had expectations of the overall performance or "quality" of a school based on the types of students serviced by the school. Parents may not, however, have had an informed view of how a school was performing relative to schools serving similar student cohorts. The similar-schools comparisons provided on *My School* are thus likely to have constituted an "information shock", to the minority of parents who looked at *My School*, and more than likely to school leaders too. Consistent with this conjecture, the most negative views of school leaders about the *My School* website were reported by schools that had the weakest performance in similar-school comparisons (Coelli, Foster & Leigh, 2018).

Table 1
School summary statistics, overall and by sector.

Statistic	Public	Independent	Catholic	All
Total school enrolment	428.6	634.9***	419.0	447.5
	(323.0)	(498.8)	(310.8)	(348.3)
Combined school (%)	5.0	73.7***	5.7	12.1
Students per tested grade (2009)	65.5	59.2***	63.5	64.4
	(60.6)	(46.9)	(54.5)	(8.1)
Student / faculty ratio	15.5	13.1***	16.3***	15.4
	(6.3)	(4.1)	(2.9)	(5.6)
Total funding ($A, 2009)	10,894	13,623***	9910***	10,965
per student	(2992)	(4428)	(2465)	(3221)
Fees / contributions ($A, 2009)	358	6206***	1698***	1231
per student	(255)	(5144)	(1442)	(2496)
Parental socioeconomic	−0.241	0.811***	0.399***	0.000
status (standardised, 2010)	(0.982)	(0.896)	(0.738)	(1.000)
Aboriginal and Torres Strait Islander	5.54	1.40***	1.87***	4.35
students (%)	(8.79)	(3.50)	(4.52)	(7.87)
Students from language backgrounds	17.6	15.2***	19.8***	17.8
other than English (%) (2009)	(22.6)	(21.5)	(24.2)	(22.9)
Female students (%)	48.7	54.2***	52.5***	50.0
	(5.8)	(17.4)	(12.2)	(9.4)
Average test scores (2008/09)	−0.096	0.270***	0.155***	−0.006
(standardised, all schools)	(0.414)	(0.362)	(0.284)	(0.409)
Average test scores (2008/09)	−0.037	0.104***	0.088***	0.003
(standardised vs similar schools)	(0.253)	(0.227)	(0.210)	(0.249)
Non-participation (%)	5.61	3.79***	3.19***	4.92
	(5.43)	(6.09)	(3.32)	(5.24)
Absence (%)	3.87	2.56***	2.07***	3.36
	(4.18)	(3.22)	(2.41)	(3.86)
Exemption (%)	1.42	0.57***	0.90***	1.22
	(2.94)	(1.51)	(2.02)	(2.67)
Withdrawal (%)	0.31	0.67***	0.23	0.33
	(1.35)	(4.42)	(0.87)	(1.85)
Schools	4811	708	1462	6981

Notes: ACARA data from the My Schools website for 2008 (except where indicated) for our estimation sample of a balanced panel of schools that excludes Special and remotely located schools. Standard deviations across schools are provided in parentheses. Statistical significance of tests of differences with public schools denoted by *** at 1% level, ** at 5% level and * at 10% level.

3. Data and descriptive statistics

Our main source of data is annual school-level information for the years 2008 to 2015. This information is precisely the same information as is provided on the *My School* website. Our analysis focuses on a balanced panel of 6981 schools, excluding Special and remotely located schools, whose test scores and non-participation rates are observed in all years on *My School* from 2008 to 2015.[10]

Average school characteristics differ across school sectors, as shown in the top panel of Table 1. Independent schools charge higher fees on average than Catholic schools. Overall resources per student are highest in independent schools and lowest in Catholic schools. Combined schools – offering education over both primary and secondary levels – are more common in the independent school sector.

Regarding student demographics, independent schools serve more advantaged students as measured by parental education and occupation, while public-school students are least advantaged, on average. Public schools serve the highest percentage of Aboriginal and Torres Strait Islander students. Catholic schools enroll the highest percentage of students from a language background other than English, and independent schools enroll the lowest percentage of such students. The percentage of female students is highest in independent schools. The higher standard deviation across independent schools relative to public

[5] 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.

[6] http://www.nap.edu.au/naplan/school-support/student-participation

[7] https://qed.qld.gov.au/programsinitiatives/education/Documents/naplan-2018-parent-perceptions-report.pdf.

[8] Examples include: https://www.heraldsun.com.au/news/schools-can-cheat-naplan-exams/news-story/8160b68c1e79ce869538913e730cdad4#:~:text=The%20Herald%20Sun%20was%20told,who%20did%20not%20do%20tests; https://www.news.com.au/lifestyle/parenting/school-life/the-littleknown-naplan-fact-parents-can-withdraw-their-child-without-a-reason/news-story/159deb3f3e81cd61e4afaa98451d26d2;https://www.smh.com.au/education/naplan-2016-parent-says-school-asked-student-to-sit-out-exam-20160512-got9tu.html.

[9] See https://www.smh.com.au/education/parents-pull-children-out-of-naplan-tests-to-avoid-my-school-ranking-20110510-1ehc6.html.

[10] Special schools enrol students with disabilities, and such students are generally exempt from NAPLAN testing. *My School* only provides information on schools with at least 5 students in tested grades.

or Catholic schools in female percentage is due to the higher prevalence of single-gender schools in this sector.

Given these differences in student background by sector, it is perhaps not surprising that NAPLAN test scores are on average lowest among public schools and highest among independent schools (middle panel of Table 1).[11] Differences in average standardised test scores across school sectors are also considerably lower (though still statistically significant) when comparisons are run amongst "similar" schools.[12] Kernel density estimates of the distributions of standardised scores using both comparison groups over the 2008 and 2009 years, by sector, are shown in Fig. 1.

The bottom panel of Table 1 shows that non-participation rates were initially highest among public schools and were driven by higher absences and exemptions. Withdrawal rates were initially highest among independent schools. A larger standard deviation in withdrawal rates was evident among independent schools even in 2008, prior to *My School*.

Fig. 2 shows the evolution of non-participation rates over time, decomposed by reason. The overall rate of non-participation rose from around 5 % in 2008 and 2009 to close to 6 % in 2010, the first year of testing after the release of *My School*, and then rose even further to around 6.7 % by 2015. The main driver of the increase in non-participation over time is withdrawal. Rates of the remaining two reasons for non-participation have barely moved over the period.

The top left panel of Fig. 3 shows non-participation rates over time, with schools split into three equal-sized groups based on their average NAPLAN performance in 2008 and 2009 relative to similar schools. While non-participation rose over time for all three groups, the largest increase, with a jump in 2010, was among those schools performing least well compared to similar schools. The top right panel of Fig. 3 shows trends in withdrawal split into the same three groups. The rise in withdrawal, detectable for all schools from 2008 to 2010, was much larger among those schools that initially performed poorly. The gap continues to grow up until 2015. Corresponding graphs for absence (bottom left) and exemption (bottom right) reveal large differences by initial relative performance, but trends (if any) are similar across these three groups.

Table 2 provides further statistics about the non-participation rate and its three components from 2008 to 2015. The evident increase in both prevalence and dispersion of non-participation are both driven by withdrawal. Absence is quite common on testing day, but its prevalence remained stable over the period. Exemption increased marginally, but its dispersion across schools remained quite stable, with a consistent 56−58 % of schools exempting no students each year. By contrast, both the prevalence and dispersion of withdrawal increased considerably. The share of schools withdrawing no students fell from 83.5 % in 2008 to 44.4 % in 2015.

4. Estimation strategy and identification

We estimate schools' responses to *My School* using the following main equation:

$$y_{it} = \gamma_i + \sum_{t=2008\ t\neq 2009}^{2015} [\alpha_t \cdot \mu_t + \beta_t (ss_i \cdot \mu_t)] + \sum_{j=1}^{4} \left[\sum_{t=2008\ t\neq 2009}^{2015} \delta_t^j \left(as_i^j \cdot \mu_t \right) \right] + \varepsilon_{it} \tag{1}$$

Here, y_{it} is the average percentage of enrolled students in tested cohorts who do not participate in testing in school i in year t; ss_i and as_i are the differences between school i's mean standardised test score[13] averaged over 2008 and 2009 and the mean standardised test score of "similar" and all schools respectively; γ_i are school indicators; μ_t are year indicators; and ε_{it} are random errors.

The β_t coefficients are of primary interest, as they capture differences in non-participation rates (overall and by reason) each year from 2008 to 2015, relative to 2009 (the omitted year), as a function of schools' initial test scores relative to similar schools. These similar school comparisons were likely the main "shock" to schools and their leaders once released on *My School*.[14] We control for a flexible fourth-order polynomial in standardised test scores relative to all other schools (as_i) to allow for the potential influence of individual student poor performance on non-participation (the second large term in Eq. (1)). These controls are also interacted with year indicators, allowing for differential trends over time in non-participation rates based on absolute initial school performance. Low-performing students are potentially less likely to participate in testing for reasons unrelated to the potential school "gaming" we are interested in understanding. By flexibly controlling for the effect of overall initial poor performance on trends in non-participation, we can isolate any effect of the potential "shock" of the similar-school comparisons.[15]

While initial test score comparisons with all and similar schools are correlated, there remains significant variation in similar school comparisons among schools grouped by overall test scores. Fig. 4 shows boxplots for each decile of initial test scores relative to all schools that reveal higher similar school score comparisons for schools with higher overall scores, as we might expect. There remains, however, significant overlap in the distributions of similar scores comparisons within overall score groups, providing the variation we rely upon during estimation.

As school-specific fixed effects are included in Eq. (1), the relationship between initial similar scores and non-participation rates is not identified.[16] Our model exploits variation over time within schools, allowing us to observe if there are changes over time in the relation between non-participation rates and initial similar score comparisons. Of most interest, we are looking for whether this relationship changes after the release of *My School*.

By focussing on a balanced panel of 6981 schools over all 8 years, we exclude 441 schools (5.9 %) observed on *My School* at website launch (2008 and 2009 data available). Of these schools, at most 129 had closed

[11] These average standardised scores are constructed by first standardising individual school average test scores separately by testing domain, grade and year. Australia-wide student-level means and standard deviations by domain, grade and cohort are employed to construct the standardisations.

[12] The Australia-wide student-level standard deviations are also employed when constructing the similar-school average standardised scores.

[13] Average percentages and test scores are calculated across all NAPLAN domains and grades tested in the school.

[14] While test score comparisons to all schools may also have been a "shock" to many parents, in this study we are primarily interested in identifying the gaming responses of schools.

[15] It is possible that schools may also respond to poor initial performance relative to all schools via "gaming". By including all-school score comparisons as controls, we may well be estimating a lower bound of the overall "gaming" effect. Our estimates may thus be conservative.

[16] Note, however, that the top left panel of Fig. 3 reveals a negative relation between non-participation and initial test score performance prior to *My School* (in 2008 and 2009), a finding confirmed by regression analysis. This is driven mostly by absences, with a minor role also played by exemptions. No relationship between initial test score performance and withdrawn percentages is evident either in the top right panel of Fig. 3 (in 2008 or 2009), or in regression analysis.

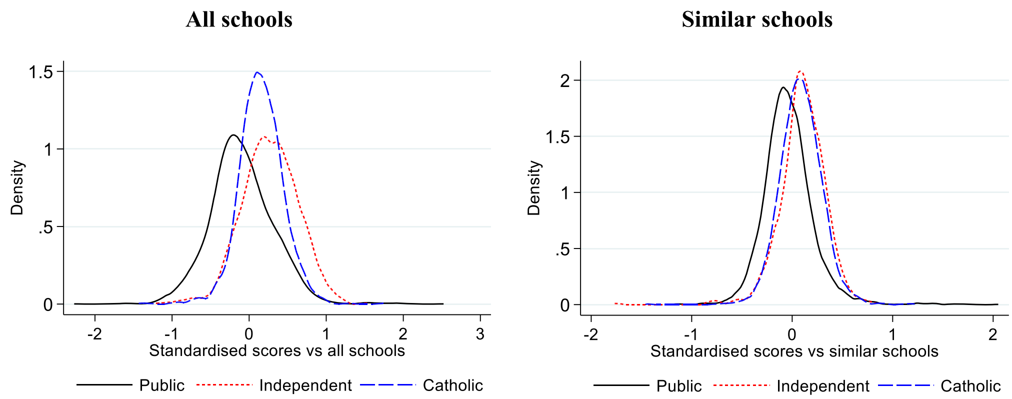


Fig. 1. Distribution of school NAPLAN scores over 2008 and 2009 by sector
Notes: ACARA data for schools observed from 2008 to 2015, excluding remote and Special schools.

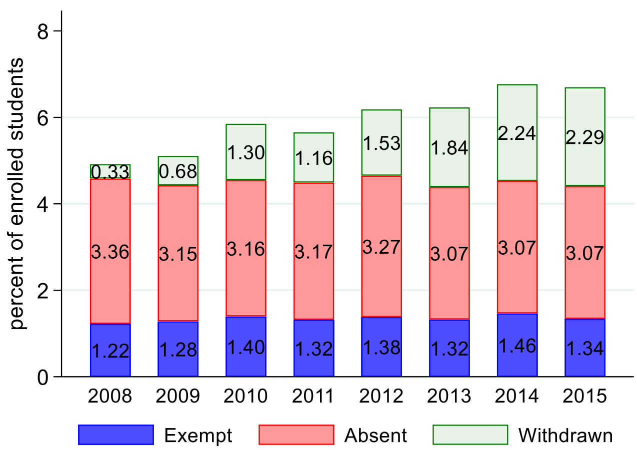


Fig. 2. Non-participation in NAPLAN testing by source and year (%)
Notes: ACARA data for 6981 schools observed from 2008 to 2015, excluding remote and Special schools.

or merged by 2015. The remaining "missing" schools are small, dropping below the threshold size for reporting on *My School* during some years. Given the very small amount of sample selection, any potential bias in our estimates is likely to be minimal.

The specific accountability intervention we study, and our estimation strategy, has important implications for the interpretation of the effects we can potentially identify, and the assumptions required for such identification. Our estimation strategy is essentially a two-way fixed effects (TWFE) model, with a "treatment" that is continuous (the similar school score comparisons) and where there are no "control" schools observed post-treatment that remain untreated (i.e., unobserved on *My School*).

Recent work by Callaway, Goodman-Bacon and Sant'Anna (2021) provides important insights into what can potentially be identified in our case and the assumptions required to do so. The authors delineate two effects that could potentially be identified in a multi-valued treatment setting. One is the *treatment effect* of some "dose" or treatment level (d), which "equals the difference between a unit's potential outcome under treatment d and its untreated potential outcome". This can be interpreted as the "level" effect of treatment: what happens if treated with dose d relative to not being treated at all. The second is the *causal response*: the effect of an incremental change in the dose at d (or the "slope" effect).

Given we do not observe untreated schools after the publication of *My School*, we cannot identify *treatment effects* as defined in Callaway et al. (2021). We can potentially identify *causal responses*, but this requires invoking what Callaway et al. (2021) call a "strong" parallel trends assumption. It is "strong" in the sense that it is not simply based on potential outcomes without treatment, equivalence of which across groups is all that is required to identify treatment effects in standard difference-in-differences designs. The "strong" version requires that the average change in potential outcomes across all units from pre-treatment to receiving some dose level d is equal to the average of the same change among those specific units that actually experienced dose d. This "strong" assumption is thus not based solely on potential outcomes without treatment. It requires a specific type of homogeneity in treatment effects, where the average effect of treatment among those units experiencing some dose d is equal to the (hypothetical) overall average effect across all units if they all had experienced dose d. It essentially rules out self-selection into dose levels. That is, it rules out units choosing a specific dose level based on their own potentially heterogeneous gains from treatment.

While this assumption may be stronger than standard parallel trends assumptions, we contend that it is reasonable in our setting. Schools do not have the ability to "choose" their similar school score comparisons. They have no role in selecting the "similar" schools used in the calculation (ACARA does this based on school ICSEA scores), and they are unlikely to be able to influence the test score outcomes of these "similar" schools.

Callaway et al. (2021) also provide insights into the specific weighted average of potentially heterogeneous causal responses across dose levels d that the TWFE estimator produces. This estimator puts more weight on units that receive doses near the mean dose $E[d]$ and the weights are "hump-shaped". This weighting may thus potentially yield an estimate that differs from the average causal response across all units in our setting if the doses observed in our data are not also "hump-shaped". The right-hand panel in Fig. 1, however, shows that our "doses" of similar school score comparisons are close to normally distributed. The TWFE estimator we employ may thus provide a reasonable approximation to an average causal response.

5. Results

5.1. Main estimates, all schools

The main results from estimating Eq. (1) for test non-participation overall and separately by reason are plotted in Fig. 5 (also reported in Appendix Table A1). We focus on the test score comparisons relative to similar schools: the "shock" in our setting. From 2010 onwards, we see negative coefficients on the year-by-initial-score interactions for non-participation overall (top left panel) and for withdrawal (top right).

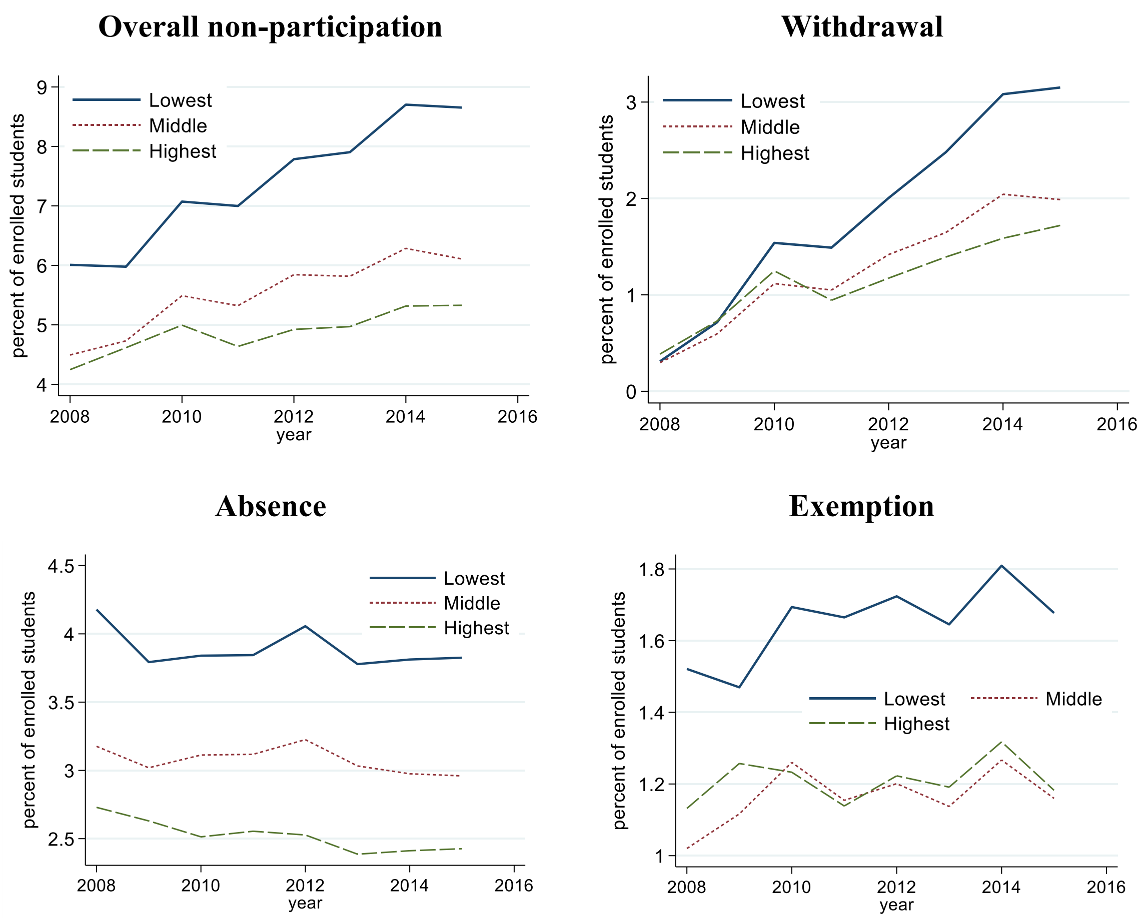


Fig. 3. Test non-participation over time by initial similar school performance
Notes: Yearly averages shown for schools in each of three equal-sized groups based on average NAPLAN test scores in 2008 and 2009 relative to similar schools.

Table 2
Distribution of non-participation in testing overall and by source over time (%).

	2008	2009	2010	2011	2012	2013	2014	2015
Non-participation								
Mean	4.92	5.11	5.85	5.65	6.18	6.23	6.77	6.70
standard deviation	5.24	5.58	6.17	5.79	6.15	6.18	6.68	6.64
percent zero	9.0	8.2	7.0	7.1	6.2	6.6	6.2	5.9
Median	3.60	3.70	4.20	4.20	4.60	4.60	5.00	5.00
75th percentile	6.60	6.70	7.80	7.40	8.30	8.45	9.10	9.00
90th percentile	10.90	10.90	12.60	11.80	13.00	13.30	14.70	14.60
Absence								
Mean	3.36	3.15	3.16	3.17	3.27	3.07	3.07	3.07
standard deviation	3.86	3.41	3.41	3.32	3.53	3.32	3.47	3.48
percent zero	13.1	13.1	12.3	12.2	11.6	12.6	13.2	12.5
Median	2.30	2.30	2.20	2.40	2.40	2.25	2.20	2.20
75th percentile	4.60	4.40	4.30	4.33	4.40	4.15	4.10	4.10
90th percentile	7.80	7.00	7.20	7.00	7.30	7.00	6.80	7.00
Exemption								
Mean	1.22	1.28	1.40	1.32	1.38	1.32	1.46	1.34
standard deviation	2.67	2.84	3.15	2.86	2.96	2.90	3.15	3.02
percent zero	55.8	55.4	55.6	56.2	55.5	57.1	55.9	58.5
75th percentile	1.50	1.50	1.50	1.50	1.50	1.50	1.50	1.50
90th percentile	3.65	4.00	4.00	4.00	4.00	4.00	4.50	4.00
Withdrawal								
Mean	0.33	0.68	1.30	1.16	1.53	1.84	2.24	2.29
standard deviation	1.84	2.91	3.50	3.24	3.69	3.97	4.33	4.40
percent zero	83.5	77.5	61.8	62.8	54.7	50.1	44.9	44.4
75th percentile	0.00	0.00	1.00	1.00	1.67	2.10	2.67	3.00
90th percentile	0.60	1.67	4.00	3.50	4.50	5.33	6.50	6.50

Notes: ACARA data for 6981 schools observed from 2008 to 2015, excluding remote and Special schools.

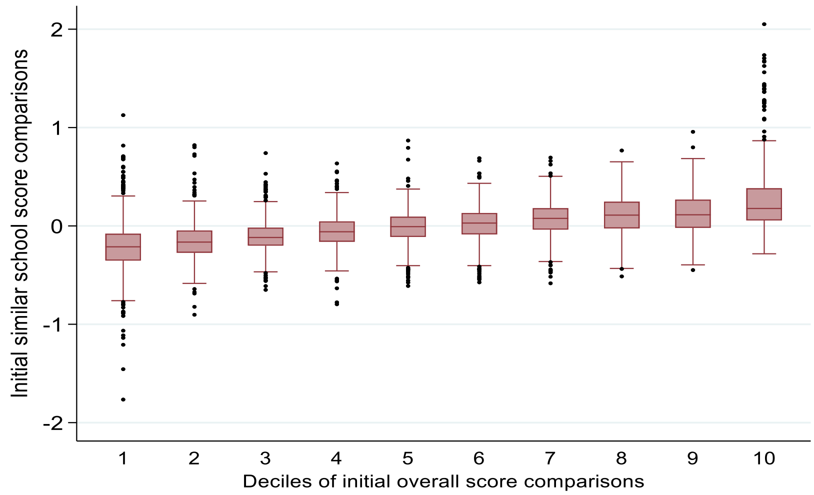


Fig. 4. Similar school comparisons within schools with comparable overall scores
Notes: ACARA data for 6981 schools observed from 2008 to 2015, excluding remote and Special schools.

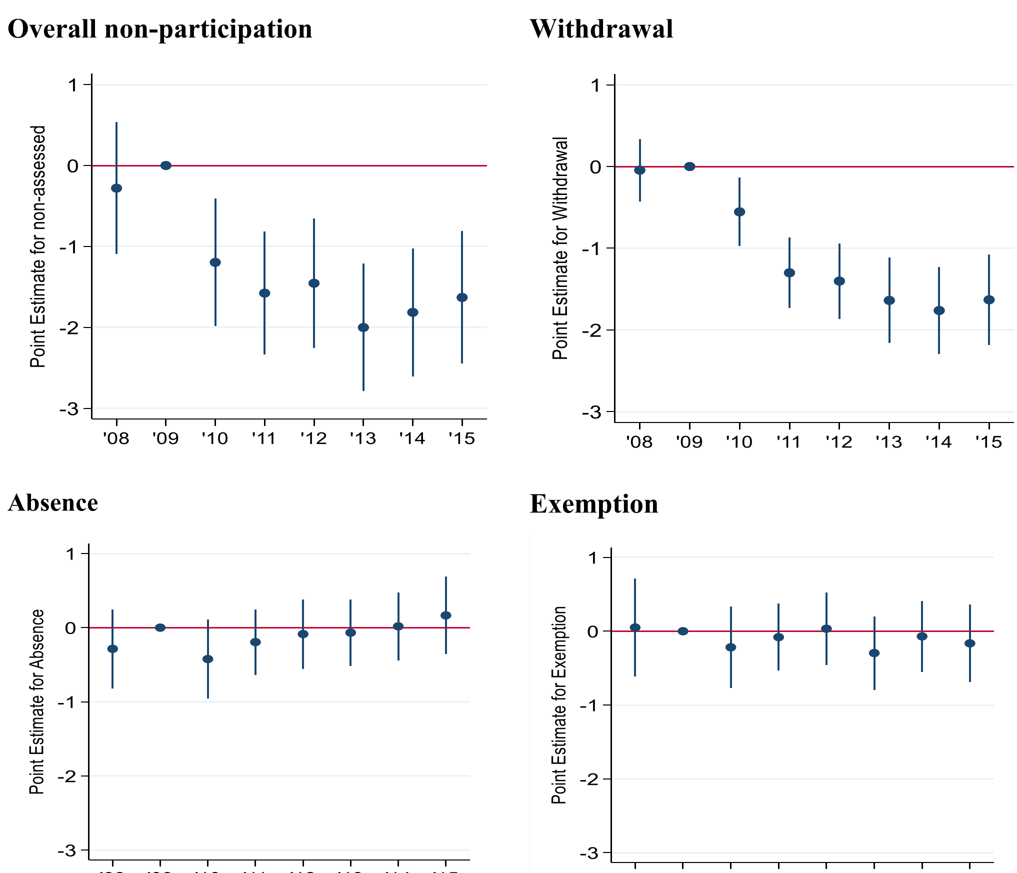


Fig. 5. Estimates of the relation each year of test non-participation (%) to initial (2008/09) school performance measured relative to similar schools
Notes: The points (lines) represent coefficient estimates (95% confidence intervals) for the beta coefficients (β_t) from Eq. (1) predicting school average test non-participation percentages as a function of test scores standardised with respect to the similar-schools average (ss_i). Estimates based on TWFE regressions (year and school) using a balanced panel of 6981 schools, excluding remote and Special schools. Regressions also include a fourth order polynomial (quartic) of average test scores standardised relative to all schools (averaged over 2008 and 2009) interacted with year indicators.

This implies that a smaller fraction of students participated in testing from 2010 onwards in Australia's worst-performing schools relative to similar schools, with this phenomenon driven by withdrawal. We see no significant relations between similar-school comparisons and changes in the remaining two reasons for non-participation: absence (bottom left panel) and exemption (bottom right).

These relationships between similar-school comparisons and withdrawal rates become increasingly more negative until 2014, with a smaller initial effect in 2010, the first year after release of *My School*. Interpreting the −1.63 coefficient on the interaction in 2015, schools with scores that are one standard deviation lower than similar schools (the standard deviation is 0.249, as shown in Table 1) have withdrawal rates that are 0.41 percentage points, or about 18 % (on the mean withdrawal rate of 2.29 in 2015, as shown in Table 2), higher than schools with average scores.

Fig. 5 shows that we also see no changes in the relationship between non-participation rates and scores relative to similar schools between 2008 and 2009, the period prior to release of *My School*. While this suggests that parallel pre-trends are evident in our setting (which is comforting), as explained above we require "strong" parallel trends for identification in this setting, a requirement that cannot be evaluated by looking at pre-trends alone.

Any of the reasons for non-participation in testing (absence, exemption or withdrawal) can potentially be manipulated by schools. Schools may catalyze learning-disability diagnoses for poorly performing students, resulting in exemption. They may also encourage parents to request withdrawal or to keep certain students at home on testing day. As illustrated in Fig. 5, increases in non-participation over the period were driven almost solely by increased withdrawal. This may be the least costly avenue to non-participation for both parents and schools. It does not appear that more poorly performing schools were enrolling a larger proportion of students with learning difficulties or who were more likely to be absent on testing day.

Our focus here is on the estimated coefficients of initial scores relative to similar schools. Test score comparisons relative to all schools are included as controls. The estimated effects of these controls at the mean level of zero are presented in Appendix Figure A1. All three components of non-participation are negatively related to initial test scores relative to all schools, with a delayed response (starting in 2012) for withdrawal and absence. These estimates may, like the estimates for similar-schools test score comparisons, reflect potential gaming, but they could also be due to changes in school composition (see Section 5) or to individual student responses to initial poor performance.

5.2. Estimates by school sector

We provide estimates broken down by school sector using the following expanded version of Eq. (1):

$$y_{it} = \gamma_i + \sum_{t=2008\ t\neq 2009}^{2015} \left[\alpha_t\cdot\mu_t + \alpha_t^I(\mu_t\cdot I_i) + \alpha_t^C(\mu_t\cdot C_i) + \beta_t(ss_i\cdot\mu_t) + \beta_t^I(ss_i\cdot\mu_t\cdot I_i) + \beta_t^C(ss_i\cdot\mu_t\cdot C_i)\right] + \sum_{K\ \in P,C,I}\left\{\sum_{j=1}^{4}\left[\sum_{t=2008\ t\neq 2009}^{2015}\delta_t^{jK}\left(as_i^{\,j}\cdot\mu_t\cdot K_i\right)\right]\right\} + \varepsilon_{it} \tag{2}$$

where P_i, I_i and C_i indicate public, independent and Catholic schools respectively. The β_t^I and β_t^C coefficients capture potential sectoral differences in effects relative to public schools. We include sector-specific unrestricted time indicators to accommodate sectoral differences in the overall trajectory of non-participation. We also control flexibly for overall test score comparison effects specific to school sector.

Estimates based on Eq. (2) are plotted in Fig. 6 (also reported in Appendix Table A2). Here we focus on overall non-participation (top panel) and on withdrawal (bottom panel). Estimates for public schools are provided in the plots on the left and reveal estimates that are similar in size and significance to those for all schools in Fig. 5. While statistically insignificant, point estimates also show that independent schools' non-participation and withdrawal responses to poor initial performance relative to similar schools were more than twice as strong as the respective responses of public schools (right of Fig. 6). By contrast, the estimated coefficients on the interaction terms for Catholic schools are near zero and even positive in some years.

The precision of our estimates for differences by school sector is reduced by estimating separate responses for each year after *My School*. To circumvent this issue, we estimated two additional versions of Eq. (2). In the first version, initial scores relative to similar schools were interacted with an indicator for all years after *My School (*2010 to 2015). In the second version, scores were interacted with an indicator for 2010 and a separate indicator for all remaining years after *My School (*2011 to 2015). The main estimates from these regressions are presented in Table 3.

The base case in Table 3 is again public schools. In both versions, and for both overall non-participation and withdrawal, the effects of similar score comparisons are negative and statistically significant among public schools.

Relationships between withdrawal rates and scores relative to similar schools are particularly evident for independent schools in these estimates, with the difference in effects as compared to public schools now statistically significant. In total (using the estimates from column 2), independent schools that are one standard deviation lower in terms of initial performance relative to similar schools have on average a 0.9 percentage point, or 39 %,[17] higher withdrawal rate. This is a sizable effect and is consistent with the large increase in the dispersion of withdrawal rates across schools reported at the bottom of Table 2.

By contrast, the estimates for the Catholic school interactions are positive, partially offsetting the negative effect estimated for public schools. Any "gaming" of the accountability system that schools may be engaging in appears less prevalent among Catholic schools.

5.3. Threats to identification

The main threat to identifying causal effects in our setting is that time-varying factors, apart from the launch of *My School,* may drive non-participation in NAPLAN testing and be correlated with initial test score performance relative to similar schools. Only time-varying factors are of concern, as any fixed factor will be absorbed by the school indicators. It is not clear *a priori* what such factors might be, but we can search for empirical indicators of them.

First, we look for signals of changes in the student population served by poorly performing schools. If such schools lose enrolments of higher-performing students, such schools will have higher proportions of students remaining who are legitimately exempted from testing, such as learning-disabled students or students from non-English speaking backgrounds, and/or students whose parents elect to withdraw them from testing for other reasons.

In Fig. 7 we plot estimates of the relations between initial performance and trajectories of overall (logged) school enrolment using Eq. (1). Changes in enrolments are positively related to our 2008/09 measures of initial performance relative to all schools (as_i, the "controls" in our main analysis), implying slower enrolment growth among the least advantaged and lowest achieving schools. For clarity, we plot coefficients on the level of as_i only. To interpret these coefficients: a school

[17] Based on the mean withdrawal rate in 2015 of 2.29 percentage points – see Table 2.

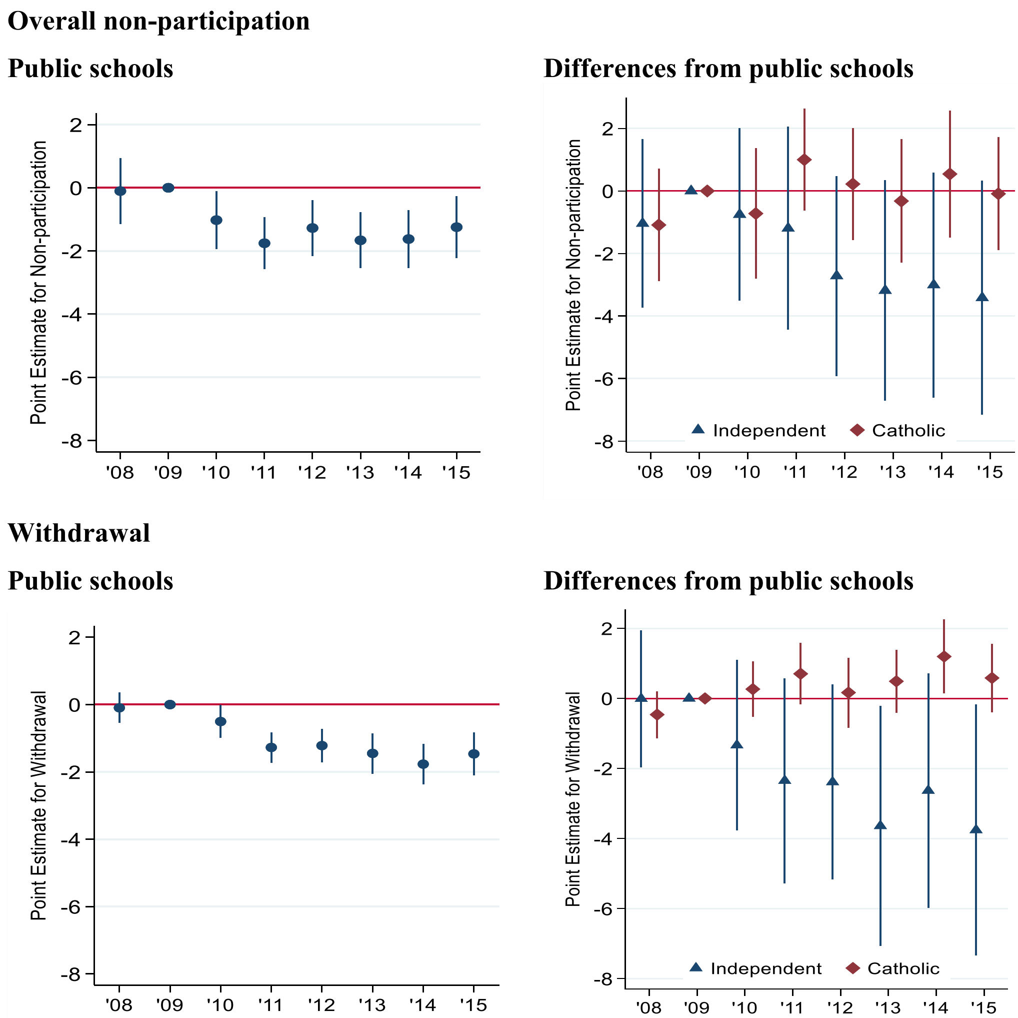


Fig. 6. Estimates of associations of non-participation in testing (%)with initial (2008/09) school performance relative to similar schools over time, by sector
Notes: The points (lines) represent the coefficient estimates (95% confidence intervals) for the beta coefficients (β_t, β_t^I and β_t^C) from Eq. (2) predicting school average test non-participation percentages as a function of test scores standardised with respect to the similar-schools average (ss_i). Estimates based on TWFE regressions (year by sector and school) using a balanced panel of 6981 schools, excluding remote and Special schools. Regressions also include a fourth order polynomial (quartic) of average test scores standardised relative to all schools (averaged over 2008 and 2009) interacted with year and school sector indicators.

that is one standard deviation higher in performance relative to all other schools (at the mean score of zero) has higher enrolment growth over 2010 to 2015 of around 0.9 percent per year.[18] Such schools thus have growth rates that are just over double the 0.7 percent annual enrolment growth of the average school. These trends are evident, however, prior to the release of *My School*. As discussed above, parents likely already held reasonably informed views prior to the release of *My School* regarding which schools are high achieving overall based on the student body they serve.

More importantly for our analysis, changes in enrolments are not related to initial performance relative to similar schools in any year we examine. It appears that parents are not reacting to this particular "information shock" by moving their children to other schools. This result is consistent with the findings of a government-commissioned review of NAPLAN and My School (Louden, 2019). Many parents of school-aged children had never looked at *My School*, and only a minority of parents had used *My School* when choosing a school for their children.[19] It is thus unlikely that student mobility can explain our estimates on test non-participation and withdrawal based on initial scores relative to similar schools.

Also, if our estimates were indeed driven by changes in the student body, we would likely see increases in all three non-participation categories – exemption, absence, and withdrawal – among poorly performing schools. As we see a response only for withdrawal, it seems unlikely that changes in the student body over time are responsible. There are differences in absences and exemptions across schools by initial performance in our setting (see Fig. 3), but these differences do not grow after the release of *My School*.

In addition, the top panel of Fig. 5 shows that the percentages of non-participating students and students withdrawn from testing in poorly performing schools began to rise significantly in the 2010 testing year. These tests were administered just four months after the *My School*

[18] This calculation is based on a predicted difference in log enrolments from 2009 to 2015 of 0.132 and a standard deviation in initial performance relative to all schools of 0.409 (see Table 1). We plot the predicted effects by year based on the full fourth-order polynomial in Appendix Figure A2. The strongest enrolment responses (i.e., the steepest predicted slopes) are at mean overall scores.

[19] Doeke (2019) also finds that parents choose schools in more complex ways than simply comparing test scores.

Table 3
Estimates of associations of test non-participation (%) with initial (2008/09) school performance measured relative to similar schools, by sector.

Non-participation	Overall (1)	Withdrawn (2)	Overall (3)	Withdrawn (4)
Score × (2010–15)	−1.380***	−1.232***		
	(0.340)	(0.201)		
Score × (2010–15) × Independent	−1.859	−2.680**		
	(1.345)	(1.249)		
Score × (2010–15) × Catholic	0.649	0.799**		
	(0.591)	(0.335)		
Score × 2010			−0.972**	−0.456*
			(0.446)	(0.254)
Score × (2011–15)			−1.461***	−1.388***
			(0.353)	(0.213)
Score × 2010 × Independent			−0.235	−1.332
			(1.267)	(1.145)
Score × (2011–15) × Independent			−2.184	−2.949**
			(1.431)	(1.338)
Score × 2010 × Catholic			−0.178	0.497
			(0.840)	(0.375)
Score × (2011–15) × Catholic			0.814	0.859**
			(0.589)	(0.357)
School indictors	✓	✓	✓	✓
Year × sector indicators	✓	✓	✓	✓
Overall score controls[a]	✓	✓	✓	✓

Notes: Estimates based on TWFE regressions using a balanced panel of 6981 schools, excluding remote and Special schools. Dependent variables are the average percentage of students in a school and year who did not participate in testing (overall and due to withdrawal) from 2008 through 2015. Initial scores are standardised relative to similar schools. Standard errors clustered at the school level are provided in parentheses. * denotes significance at 10%, ** at 5% and *** at 1% levels.
[a] Overall score controls are a fourth order polynomial (quartic) of average test scores standardised relative to all schools (averaged over 2008 and 2009) interacted with year by sector indicators.

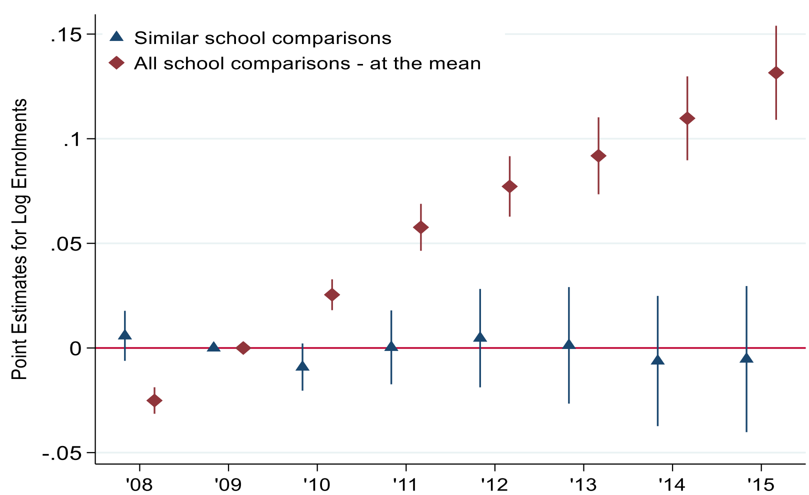


Fig. 7. Responses of school enrolment to initial (2008/09) school performance relative to both similar schools and all schools, yearly from 2008 to 2015
Notes: The points (lines) represent coefficient estimates (95% confidence intervals) from Eq. (1) predicting school log enrolments as a function of test scores standardised with respect to both similar schools (ss_i) and all schools (as_i). Estimates based on TWFE regressions (year and school) using a balanced panel of 6981 schools, excluding remote and Special schools. .

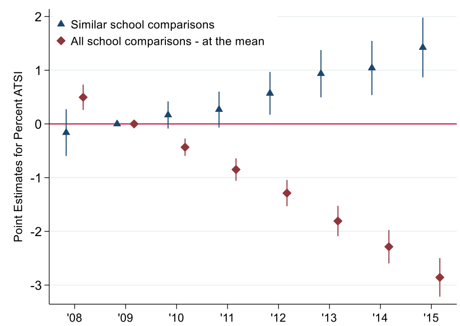
website launch in January 2010. At this point, with enrolments for the 2010 school year all but finalised, even highly motivated school leaders wanting to take action to shore up their apparent performance in the next update of *My School* would have had few options left. Removing poorly performing students from the 2010 testing pool through urging their parents to withdraw them was one of few options still available.

We delve further into potential changes in student populations over time in Fig. 8. We again estimate versions of Eq. (1), but with the dependent variables now demographic characteristics of the school population. The three demographic characteristics available to us on a consistent basis over the period are: the percentage of students who are Aboriginal and Torres Strait Islander, the percentage of students from a language background other than English and the percentage of students who are female. Estimates from predicting each of these three demographic variables are presented in Fig. 8.[20]

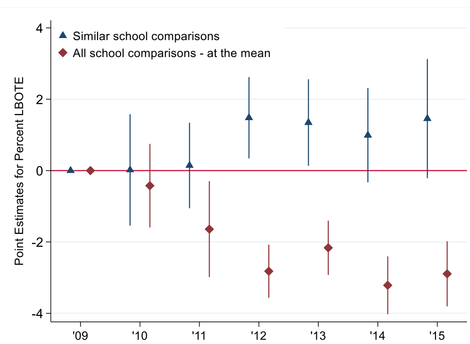
In panel A of Fig. 8, we see that the percentage of Aboriginal and Torres Strait Islander students is increasingly negatively related to initial scores relative to all schools at the mean of zero. That is, over time, such students appear on average to be increasingly more likely to be enrolled in the least advantaged and poorest performing schools. However, this share is associated increasingly positively with initial scores relative to similar schools: the source of the “shock” in our analysis. Note also that the overall percentage of students identifying as Aboriginal and Torres

[20] Due to large changes in the construction of the variable denoting parental socio-economic status over the period we examine, we do not present estimates for this demographic characteristic.

Panel A: Percent of Aboriginal and Torres Strait Islander (ATSI) students



Panel B: Percent of students from a language background other than English (LBOTE)



Panel C: Percent of female students

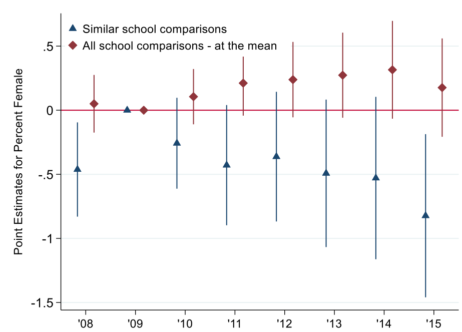


Fig. 8. Estimated relationships between school demographics with initial (2008/09) performance relative to similar and all schools over time
Notes: The points (lines) represent coefficient estimates (95% confidence intervals) from Eq. (1) predicting the three school demographic percentages as a function of test scores standardised with respect to both similar schools (ss_i) and all schools (as_i). Estimates based on TWFE regressions (year and school) using a balanced panel of 6981 schools, excluding remote and Special schools. .

Strait Islander in our balanced panel of schools grew by 2 percentage points from 2008 to 2015. It is thus highly likely that the strong negative relation with initial test scores relative to all schools largely reflects this growth, rather than differential student mobility.

Similar but less precisely estimated patterns are observed for the percentage of students from a language background other than English in Panel B. This percentage also grew by two percentage points from 2009 to 2015, so these estimates may largely reflect this growth.[21] As can be seen in Panel C of Fig. 8, the opposite pattern is observed for the female share variable, the coefficients are quite imprecisely estimated, and the effect sizes are much smaller.

Overall, there is no clear indication here that changes in student composition are likely to significantly impact our estimates on withdrawal from testing in response to poor initial performance relative to similar schools.

[21] This variable is only available from 2009 onwards.

Finally, student mobility would need to be extremely large to be the primary driver of our estimates. To demonstrate, we consider a simple yet extreme example. Take one poorly performing school that is a full one unit below zero in terms of score comparisons against all schools (i.e., 2.5 standard deviations). Take a second school that is performing one full unit above zero. The annual relative change in enrolment based on Fig. 7 between these two schools in 2010 would be around 4 percent. Suppose that this change was brought about because 2 percent of the students in the poorly performing school moved to the high-performing school. Suppose that none of these mobile students withdrew from testing in their new school, and that their new school had no withdrawal prior to *My School* (most schools had zero withdrawal). Further suppose that the withdrawal rate among students in the poorly performing school was 5 percentage points before *My School* (placing it above the 95th percentile in terms of withdrawal rates). The loss of 2 percent of its students would raise the withdrawal rate among remaining students at the poorly performing school by just 0.1 percentage point. This is an order of magnitude lower than the 1.1 percentage point effect we

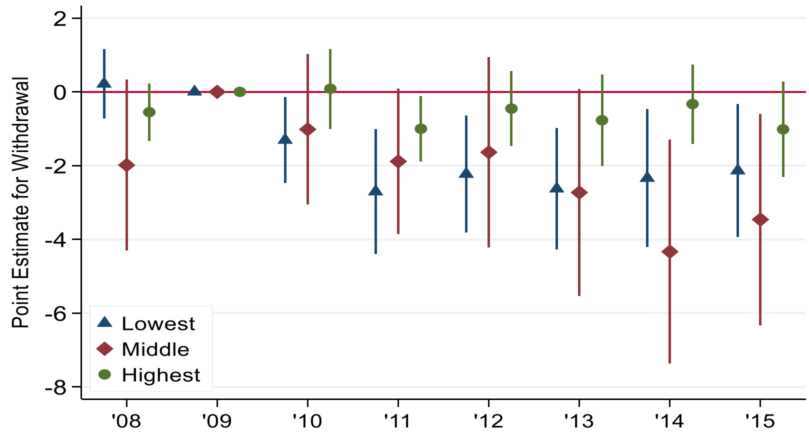


Fig. 9. Estimates from predicting withdrawal from testing (%) based on initial (2008/09) school performance relative to similar schools over time, by tercile
Notes: The points (lines) represent coefficient estimates (95% confidence intervals) for the beta coefficients (β_t) from an expanded version of Eq. (1) predicting school average withdrawal from testing percentages as a function of test scores standardised with respect to similar schools (ss_i). Separate linear effects are estimated by tercile of ss_i. Estimates based on TWFE regressions (year by tercile and school) using a balanced panel of 6981 schools, excluding remote and Special schools. Regressions also include a fourth order polynomial (quartic) of average test scores standardised relative to all schools interacted with year indicators. .

Table 4
Estimates of potentially non-linear relations of withdrawal from testing (%) to initial (2008/09) school performance measured relative to similar schools.

	Piecewise linear	Fourth order polynomial
Initial score × (2010–2015) × lowest tercile	−2.332***	
	(0.666)	
Initial score × (2010–2015) × middle tercile	−1.519*	
	(0.789)	
Initial score × (2010–2015) × highest tercile	−0.307	
	(0.390)	
Initial score × (2010–2015)		−1.519***
		(0.206)
Initial score squared × (2010–2015)		1.900***
		(0.515)
Initial score cubed × (2010–2015)		0.236
		(0.358)
Initial score quartic × (2010–2015)		−0.567**
		(0.236)
School indictors	✓	✓
Year indicators	✓	✓
Overall score controls[a]	✓	✓

Notes: Estimates based on TWFE regressions using a balanced panel of 6981 schools, excluding remote and Special schools. The dependent variable is the average percentage of students in a school and year who were withdrawn from testing. Initial scores are standardised relative to similar schools. Standard errors clustered at the school level are provided in parentheses. * denotes significance at 10%, ** at 5% and *** at 1% levels.
[a] Overall score controls are a fourth order polynomial (quartic) of average test scores standardised relative to all schools interacted with year indicators.

estimate for 2010, an effect that grows to 2.6 percentage points in 2011 (Appendix Table A1).

As noted in Section 4, identification of causal responses in our setting requires a "strong" version of parallel trends to hold. We believe that such a requirement is likely to hold in our setting. Callaway et al. (2021) suggest that such a requirement may be more defensible when considering units (schools in our setting) that have similar "doses" of treatment (i.e., similar initial scores relative to similar schools). This suggests estimation focusing on schools with initial scores that are close to each other, or in other words constructing "local" estimates, which may potentially reveal responses that are not necessarily linear.

Following this logic, we present estimates allowing for potential non-linearities in withdrawal responses to poor initial performance relative to similar schools in several ways. To begin, in Fig. 9, we plot estimated effects for withdrawal by year that can differ by tercile of initial scores relative to similar schools. This is essentially a piecewise linear model, allowing separate linear effects within each tercile by year. While these estimates are imprecise, we can see more negative effects for the lowest and middle terciles, while effects for the highest tercile are closer to zero. This makes intuitive sense, as performance below similar schools is likely to raise concerns among school leaders and motivate potential manipulation of the testing pool, while performance above similar schools is less likely to raise such concerns.

To increase the precision of estimates while also allowing for potential non-linearities, we next turn to estimating one effect for all years post *My School (*2010 to 2015). We focus on the effects on withdrawal, estimating alternately a piece-wise linear model and a model that includes a 4th order polynomial of scores compared to similar schools. The estimates are reported in Table 4. Predictions from the polynomial model are depicted in Fig. 10. There are clear non-linear effects. Effects are large among the weakest-performing schools, while effects are near zero among above-average schools.

Finally, could our results merely be reflecting parents' responses to an increased focus on test performance? If schools with poor initial scores subsequently put more emphasis on students' NAPLAN test performance, their students may have become more anxious, leading to a spike in withdrawals as parents responded to their children's anxieties by withdrawing them from testing. This mechanism is plausible and would not have logically affected exemption or absence rates. Yet even if parents rather than schools drove these withdrawal responses, there remain two unintended consequences of increased withdrawal that warrant attention: (1) certain students are not being tested, so schools and the education system more widely are less informed regarding which students need further support; and (2) the informational content of school average test scores is reduced, clouding the information available to policy-makers about where there is greatest need for assistance, and to parents about the performance of different schools.

5.4. Non-participating students

The estimates presented above indicate that poorly performing schools may attempt to raise their displayed *My School* NAPLAN performance by manipulating the pool of students being tested, via encouraging parents of low-performing students to withdraw their children from testing. As noted above, withdrawal is formalised by parents informing the school principal of their wishes "in the manner specified by the local testing authority", usually by filling out a simple form. A natural question is to see what each year's pool of withdrawn

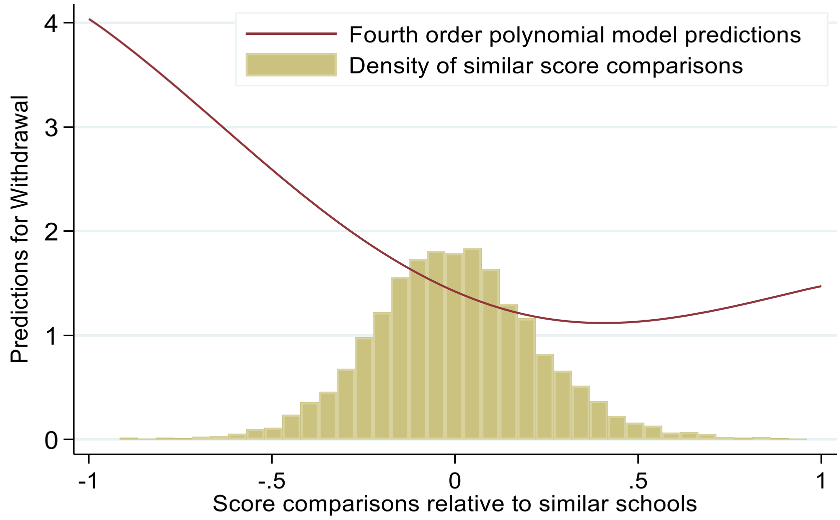


Fig. 10. Polynomial model estimates of the association between withdrawal from testing (%) and initial (2008/09) school performance relative to similar schools
Notes: Estimates based on TWFE regressions (year and school) using a balanced panel of 6981 schools, excluding remote and Special schools. Regressions also include a fourth order polynomial (quartic) of average test scores standardised relative to all schools interacted with year indicators. .

Table 5
Individual student achievement in 2011 predicted by 2013 test non-participation status.

	Standardised	Proportion	Any in	School rank
	score	**lowest band**	**lowest band**	**(0 to 1 = top)**
	(1)	(2)	(3)	(4)
Absent	−0.442***	0.0537***	0.126***	−0.113***
	(0.0134)	(0.0028)	(0.0061)	(0.0038)
Withdrawn	−1.095***	0.203***	0.430***	−0.308***
	(0.0249)	(0.0079)	(0.0138)	(0.0065)
Exempt	−1.403***	0.305***	0.579***	−0.371***
	(0.0340)	(0.0138)	(0.0211)	(0.0079)
Private school	0.226***	−0.0232***	−0.0614***	−0.00459***
	(0.0119)	(0.0009)	(0.0024)	(0.0004)
Absent × Private school	0.150***	−0.0286***	−0.0609***	0.0375***
	(0.0246)	(0.0044)	(0.0101)	(0.0078)
Withdrawn × Private school	0.0511	−0.0646***	−0.0823**	0.00155
	(0.0570)	(0.0149)	(0.0321)	(0.0165)
Exempt × Private school	−0.0163	−0.0644**	−0.0767	−0.0117
	(0.0732)	(0.0287)	(0.0477)	(0.0172)
Constant	−0.0373***	0.0403***	0.121***	0.522***
	(0.0079)	(0.0007)	(0.0018)	(0.0003)
Students	230,584	230,584	230,584	230,584
R-squared	0.047	0.053	0.039	0.018

Notes: ACARA data for students in grade 5 in 2013 with observed 2011 (grade 3) NAPLAN test scores, excluding students in remote schools, in schools with a high proportion of Exempt students (likely Special schools) and in schools with less than 5 students in the tested grade. Robust standard errors clustered at the school level are provided in parentheses. * denotes significance at 10%, ** at 5% and *** at 1% levels.

students looks like: would their NAPLAN scores in fact be likely to bring down schools' average scores?

Unfortunately, we are unable to access individual student-level data for the period covering the initial launch of the *My School* website in 2010. We do, however, have access to anonymised student-level data from 2013 onwards, including the NAPLAN test results of these students (where available) in the prior testing year. We use these data to evaluate whether students who did not participate in testing in 2013 were poor performers when tested in 2011. We focus on students expected to sit the grade 5 tests in 2013, for whom we also have (if available) the scores they received on NAPLAN tests taken in grade 3, in 2011.

Table 5 shows estimates from regressions of four alternative individual grade 3 student achievement measures for 2011 on indicators for withdrawal, absence or exemption from testing in 2013, a private school indicator,[22] plus interactions of the private school indicator with the withdrawal, absence and exemption indicators. In column 1, the dependent variable is the average across the 5 testing domains of the standardised scores[23] of each student who sat grade 3 NAPLAN tests in 2011.

Public school students absent on testing days in 2013 scored 0.442 of a standard deviation below average in 2011, while students withdrawn from testing in 2013 scored over a full standard deviation below average, and exempt students scored 1.4 standard deviations below average. The interactions of the private school indicator (set to 1 for both independent and Catholic schools) and each of the three forms of non-participation indicate that the weakest students were also being withdrawn or exempted from testing in non-public schools, but to no more of a degree than in public schools.

ACARA categorises all student NAPLAN test scores into one of ten bands, and we can use data on these bands to further interrogate the relationship between student performance in 2011 and subsequent participation in testing in 2013. Grade 3 students' scores range across the lowest six bands, grade 5 scores range from band 3 to band 8, grade 7 scores range from band 4 to band 9 and grade 9 scores range from band 5 to band 10. Thus, each grade has scores that span precisely 6 bands. We construct indicators for students who scored in the lowest band applicable to their specific school grade (i.e., bands 1, 3, 4 and 5 for grades 3, 5, 7 and 9, respectively). These students would be considered to have performed below expectations. On average across our balanced panel of schools, 6 % to 7 % of students scored in the lowest band relevant for their grade.

In column 2 of Table 5, the dependent variable is the proportion of the (up to five) tests taken in 2011 on which the student scored in the lowest band. In column 3, the dependent variable is an indicator for whether the student scored in the lowest band on any of the tests taken in 2011. Using either measure, we see that test non-participants are

[22] It was not possible to separately identify independent from Catholic private school students in the individual student datasets.

[23] Scores were standardised to have a mean of zero and standard deviation of 1 across all students sitting each test across all schools in 2011.

more likely to be weaker students. Yet again, exempt students are the weakest, followed by withdrawn students, and then absent students, and there is no evidence that private school non-participants are more negatively selected than public school students.

Finally, in column 4 of Table 5, the dependent variable is the 2011 within school-by-cohort rank of the student based on his or her grade 3 standardised test scores. We scale these ranks for comparability across schools of different sizes such that the weakest student in a school and cohort has rank $1/n$, where n is the size of the 2011 school cohort sitting the grade 3 tests. The strongest student has rank equal to 1 (i.e., n/n). By predicting this measure based on non-participation in 2013 testing, we can determine whether the weakest students in a school are being withdrawn from testing, regardless of the average performance of the school. As with the other achievement measures, the evidence in column 4 suggests that exempt students occupy the bottom of the school distribution, while withdrawn students do not place much higher.

Student mobility across schools in Australia is not uncommon. However, this mobility is more often related to disruption and poorer student performance (Lu & Rickard, 2016) than to purposeful changes to better-performing schools. The individual-level student data we have access to also has an indicator for whether a student was at the same school when tests were taken two years prior. For students in grade 5 in 2013, around 17 % were enrolled in a different school when they sat NAPLAN tests two years prior when in grade 3. Mobile students were, however, less likely to participate in testing in their new school (by around 1.6 percentage points). All three reasons for non-participation were higher among mobile students. Mobile students are more likely to be withdrawn, absent and exempt from testing by around 0.5, 0.8 and 0.3 percentage points respectively.

We observed differential trends in some student demographics related to initial test scores in Fig. 8. The individual-level student data can also be used to understand whether withdrawal rates differ by demographic group. For students in grade 5 in 2013, Aboriginal and Torres Strait Islander students were a statistically significant 0.8 percentage points *more* likely to be withdrawn from testing than other students, but this difference evaporates once we condition on student scores in grade 3 in 2011.[24] Students from a language background other than English (LBOTE) were *less* likely to be withdrawn from testing than other students by a statistically significant 0.5 of a percentage point.[25] While this difference declines by one third once we condition on prior test scores, it remains statistically significant. Female students have a withdrawal rate that is lower than males by 0.3 percentage points, with this difference declining by two-thirds once we condition on prior test scores. These observed differences by demographic group are unlikely to undermine our main estimates on withdrawal.[26]

While the estimates in Table 5 clearly establish that students withdrawn from testing in 2013 were generally weaker NAPLAN performers in 2011, these results do not tell us about responses over time after the launch of *My School*, nor about whether responses were stronger amongst schools that performed poorly in 2008 and 2009.

In our final empirical tests, we ask whether schools whose non-participation rates increased the most also had larger reductions in the percentage of students receiving low scores. Such a pattern – presumably the effect intended by schools whose leaders encouraged the parents of poorly performing students to withdraw them from NAPLAN testing – could be interpreted as evidence of this "gaming" response to the launch of *My School*.

Table 6
Estimates of relations between proportions of students scoring in the lowest band and test non-participation.

	Change percent lowest band 2008/09 to 2011/12			
	(1)	(2)	(3)	(4)
Change non-participation% 2008/09 to 2011/12	−0.0957***	−0.116***		
	(0.0086)	(0.0088)		
Change withdrawn% 2008/09 to 2011/12			−0.144***	−0.162***
			(0.0133)	(0.0136)
percent in lowest band in 2008/09	−0.110***		−0.114***	
	(0.0058)		(0.0057)	
R-squared	0.073	0.024	0.072	0.020
	Change percent lowest band 2008/09 to 2013/15			
Change non-participation% 2008/09 to 2013/15	−0.0703***	−0.118***		
	(0.0085)	(0.0089)		
Change withdrawn% 2008/09 to 2013/15			−0.156***	−0.204***
			(0.0115)	(0.0120)
percent in lowest band in 2008/09	−0.189***		−0.187***	
	(0.0065)		(0.0064)	
R-squared	0.131	0.025	0.145	0.040
Schools	6981	6981	6981	6981

Notes: Dependent variable is the change in the percent of students scoring in the lowest of the six bands relevant for their specific grade in 2012. Robust standard errors are provided in parentheses. * denotes significance at 10%, ** at 5% and *** at 1% levels.

In Table 6, we present the results of simple regressions of changes in the percentage of students in a school who scored in the lowest band (as described above) on changes in the percentage of students in that school who did not participate in testing, overall and via withdrawal. In our preferred estimates (columns 1 and 3), we also include a control for the initial percentage of students at the school who scored in the lowest band in 2008/09. This controls for mean reversion in student outcomes over time.

In Columns 1 and 2 of Table 6, the focal independent variable is the change in any form of non-participation, while in Columns 3 and 4 it is the change only in the percentage of students withdrawn from testing. In both cases, and over both time periods, a significant negative association is found, with the size of the relationships up to twice as large when focusing on withdrawn percentages only. If a school's withdrawn percentage increased by one percentage point from 2008/09 to 2011/12 (the "initial" response), the percentage of its students scoring in the lowest band is estimated to have fallen by 0.144 percentage points on average, holding constant the percent of its students scoring in this lowest band in 2008/09.

The results in Table 6 are suggestive of the selective withdrawal of the worst performing students after the release of *My School*. Although it is still possible that this association is due to another, more subtle cause – such as parents of low-performing students enrolled in poorly performing schools disproportionately deciding that testing is bad for their children, and on that basis electing to withdraw them – we view our findings as most consistent with schools having encouraged more low-performing students not to sit the NAPLAN tests in response to the accountability shock embodied in *My School*.

6. Conclusions

We find evidence that Australian schools responded predictably to the increase in accountability ushered in by the *My School* website in 2010. The response of assessing fewer poorly performing students arguably works against the objectives of increased accountability. Such behavior shields poorly performing schools, while under-providing

[24] This group was also more likely to be absent on testing day or to be exempt from testing, by 2.2 and 0.6 percentage points respectively.

[25] This group was also less likely to be absent on testing day or to be exempt from testing, by 0.7 and 0.1 percentage points respectively.

[26] Even the predicted growth in LBOTE of 1.5 percentage points from 2009 to 2015 due to a one standard deviation higher performance relative to similar schools, coupled with a 0.5 percentage point lower withdrawal percentage, yields a lower overall withdrawn percentage of just 0.0075. This is a fraction of the negative 1.6 percentage point estimate of Fig. 5 and Table A1 for 2015.

performance information to Australia's weaker students and those who provide for them. We find some evidence that those schools with potentially the most to lose from public accountability – private schools – respond more strongly than public or, especially, Catholic schools by adjusting their testing pool.

Australia is not the first country where this type of response has been observed (Cullen & Reback, 2006; Figlio & Getzler, 2006). Possible policy remedies could include requirements that a minimum percentage of students in each school participates in testing each year; that schools explain any significant year-over-year changes in non-participation; and/or that schools that report large drops in test participation are automatically audited by the government. Without such safety nets, well-intentioned reforms towards improving school accountability may only make things worse for Australia's most vulnerable students.

Of course, schools may have initiated responses to *My School* that have positive impacts on poorly performing students, such as devoting more resources to teaching, streaming classes, mandating more class time on core subjects, instituting smaller classes, and so forth. In another recent paper (Coelli et al., 2018) we look for evidence of these types of responses using survey data from school principals from before and after the *My School* launch. We found little consistent evidence of changes that we might expect to improve outcomes of poorly performing students in those survey responses.

Our results, while striking, still leave open the question of what precisely drove schools' gaming responses in Australia. Do schools believe parents will respond to school performance data published on *My School* via changing their enrolment decisions? It does not appear that parents have responded that way. Are responses driven by a desire to "keep parents happy" or to underpin public perceptions of the "prestige" of the school? It is unlikely that schools are responding due to fearing consequences from the government or some other oversight body, as those bodies already had access to NAPLAN results prior to *My School*. Isolating the main drivers of gaming responses to the accountability mechanism in the Australian setting is a fruitful area for future research.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

We thank Mark Wooden, three anonymous referees and Celeste Caruthers (Editor) for helpful comments and suggestions, in addition to seminar participants at the University of New South Wales, University of Melbourne, SOLE 2015, AASLE 2019, and ACE 2021 meetings. We are grateful to Jenny Donovan at the center for Education Statistics and Evaluation in the NSW Office of Education for invaluable background information, and to Van Dao, Narin Chhun, Sophie Fraser and the rest of the ACARA team for their support in extracting and explaining the NAPLAN data. We are greatly indebted to David Figlio and Andrew Leigh for their efforts in initiating the broad project and making significant contributions during the development of this paper. We also thank Gordon Leslie and Su-Min Lim for superior research assistance, and the University of New South Wales and the University of Melbourne for financial support. All views expressed and any remaining errors are ours, and should not be attributed to ACARA, the New South Wales Department of Education, or any other party.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.econedurev.2024.102523.

References

Andrabi, T., Das, J., & Khwaja, A. I. (2017). Report cards: The impact of providing school and child test scores on educational markets. *American Economic Review, 107*(6), 1535–1563. https://doi.org/10.1257/aer.20140774

Callaway, B., Goodman-Bacon, A., & Sant'Anna, P. (2021). Difference-in-differences with a continuous treatment. *Working paper*. https://doi.org/10.48550/arXiv.2107.02637. arXiv:2107.02637v2.

Camargo, B., Camelo, R., Firpo, S., & Ponczek, V. (2018). Information, market incentives, and student performance: Evidence from a regression discontinuity design in Brazil. *Journal of Human Resources, 53*(2), 414–444. https://doi.org/10.3368/jhr.53.2.0115-6868R1

Cilliers, J., Mbiti, I., & Zeitlin, A. (2021). Can Public rankings improve school performance? Evidence from a nationwide reform in Tanzania. *Journal of Human Resources, 56*(3), 655–685. https://doi.org/10.3368/jhr.56.3.0119-9969R1

Coelli, M., Foster, G., & Leigh, A. (2018). Do school principals respond to increased public scrutiny? New survey evidence from Australia. *Economic Record, 94*, 73–101. https://doi.org/10.1111/1475-4932.12400 [Special Issue].

Cullen, J., & Reback, R. (2006). Tinkering towards accolades: School gaming under a performance accountability system. In T. Gronberg, & D. J& ansen (Eds.), *Improving school accountability (Advances in applied microeconomics 14)* (pp. 1–34). Emerald Group Publishing Limited. https://doi.org/10.1016/S0278-0984(06)14001-8.

Doeke, E. (2019). *More than a socio-economic variable: Family strategies of educational advantage within Germany and Australia. Phd Thesis*. Melbourne Graduate School of Education, University of Melbourne. https://rest.neptune-prod.its.unimelb.edu.au/server/api/core/bitstreams/841fe955-aa94-5e86-a4ef-e3a304fb4ba6/content.

Figlio, D. (2006). Testing, crime and punishment. *Journal of Public Economics, 90*(4–5), 837–851. https://doi.org/10.1016/j.jpubeco.2005.01.003

Figlio, D., & Getzler, L. (2006). Accountability, ability and disability: Gaming the system? In T. Gronberg, & D. Jansen (Eds.), *Improving school accountability (Advances in applied microeconomics 14)* (pp. 35–49). Emerald Group Publishing Limited. https://doi.org/10.1016/S0278-0984(06)14002-X.

Figlio, D., Loeb, S., Hanushek, E., Machin, S., & Woessmann, L. (2011). School accountability. In *Handbooks in economics: Economics of education, 3* pp. 383–421). Amsterdam: North-Holland https://cepa.stanford.edu/sites/default/files/Accountability_Handbook.pdf.

Figlio, D., & Winicki, J. (2005). Food for thought: The effects of school accountability plans on school nutrition. *Journal of Public Economics, 89*(2–3), 381–394. https://doi.org/10.1016/j.jpubeco.2003.10.007

Firpo, S., Ponczek, V., & Possebom, V. A. (2014). *Private education market, information on test scores and tuition practices. iza discussion paper no. 8476*. Bonn, Germany https://www.iza.org/publications/dp/8476/private-education-market-information-on-test-scores-and-tuition-practices.

Florida Department of Education. (2014). *2014 Guide to calculating school grades*. Bureau of Accountability Reporting; Division of Accountability, Research, and Measurement. http://schoolgrades.fldoe.org/pdf/1314/SchoolGradesCalcGuide2014.pdf.

Jacob, B. A., & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *Quarterly Journal of Economics, 118*(3), 843–877. https://doi.org/10.1162/00335530360698441

Louden, W. (2019). *NAPLAN reporting review*. Prepared for COAG Education Council: Education Services Australia. https://files.eric.ed.gov/fulltext/ED599927.pdf.

Lu, L., & Rickard, K. (2016). *Mobility of students in NSW government schools*. Centre for Education Statistics and Evaluation. https://education.nsw.gov.au/about-us/education-data-and-research/cese/publications/research-reports/mobility-of-students-in-nsw-government-schools.

Mizala, A., & Urquiola, M. (2013). School markets: The impact of information approximating schools' effectiveness. *Journal of Development Economics, 103*(2013), 313–335. https://doi.org/10.1016/j.jdeveco.2013.03.003