Beyond the threshold: The implications for pupil achievement of reforming school performance metrics

Simon Burgess, University of Bristol and IZA
Dave Thomson, FFT Education Datalab
Discussion Paper 22/770

August 2022

School of Economics

University of Bristol
Priory Road Complex
Bristol
BS8 1TU
United Kingdom

University of BRISTOL
School of Economics

Beyond the threshold: The implications for pupil achievement of reforming school performance metrics

Simon Burgess, University of Bristol and IZA

Dave Thomson, FFT Education Datalab

August 2022

Abstract

We study the effects of a major change to the school accountability system in England. In 2015, the leading published school performance metric was switched from a threshold measure (essentially the fraction of students above a test score level) to an average score measure. Using 7 years of data on all secondary schools in England, we show that this intervention relatively reduced the test scores of students near the threshold, in favour of groups above the threshold (marginally) and below (substantially). We check the sensitivity of our results to different decisions, and present findings on heterogeneous treatments.

We are very grateful to the Nuffield Foundation for funding this research, and to the Department for Education for access to the National Pupil Database and to members of our Advisory Group for comments on earlier versions of this report. We particularly want to thank Ellen Greaves and Hans Sievertsen for detailed comments. Any errors are ours.

1. Introduction

Around the world, many countries have adopted school accountability systems: Bergbauer et al (2021) report that in 2015 around two thirds of schools in their sample of 59 PISA-taking countries are subject to accountability[1]. We define school accountability as the public provision of school-performance information, on a regular basis, in the same format, and using independent metrics. Many of these schemes are well-established; for example, in England the core of the system has been in place for 30 years now, and in the US, for around 20 years. Overall, the evidence suggests that public accountability for schools has raised pupil achievement, though researchers have also provided evidence of gaming, behaviour distortion and some cheating (see reviews Figlio and Loeb, 2011; Dee, 2020, Burgess and Greaves, 2021; and international comparisons in Woessmann, 2016).

The impact of accountability on a school system depends crucially on its design, including the nature of the metrics targeted, the measurement system, the implications of failing to meet targets and more. Here, there is less evidence on the implications of the detailed design of accountability systems for pupil achievement. This is the contribution of this paper. We study a substantive change in the metric used in school performance tables in England and analyse the causal impact on the distribution of test scores. Crucially, other aspects of the accountability remained the same. In 2013, the UK Government announced a reform to the accountability framework for state-funded schools in England, proposing to publish a new performance indicator, Progress 8. There are two key facets of the change. First, the focus here, is that the calculation of school performance switched from being a threshold measure to being a simple average. In the prior regime, the key metric was the fraction of pupils achieving a benchmark exam outcome; in the new regime it is the average over pupils. Second, the new headline figure is a value-added measure as opposed to a simple raw outcome figure. This may well have important effects on the schools' admission priorities – leading schools to prefer to admit students thought to be high value-added rather than simply high raw score pupils. However, to focus our attention on the first aspect, we side-step admissions change by using data on pupils already assigned to their school at the time of the reform.

The switch from a threshold design to a simple average frees Headteachers (HTs) to choose groups of pupils to focus discretionary resources on, or equally on all. We sketch below a framework for

[1] Using their variable "School-focused external comparison" (question: "In your school, are assessments of 15-year-old students used for any of the following purposes?" Answer: "To compare the school to district or national performance?"); in table A1 the average of country averages is 66%. In fact, each individual country average is greater than 0 (minimum 17%), so arguably is accountability even higher than that.

understanding this decision. Since there is little prior knowledge on the underlying empirical functions, we are agnostic about what might result and adopt a flexible approach to the analysis. The (universal) policy change gives us one difference; the fact that the policy will have its most intense effects on a particular group of pupils (those near the performance threshold) gives us the other. We use a difference-in-differences model, exploiting the introduction of Progress 8 and we isolate the key group of pupils using only pre-reform variables. This policy change happens at one date across all secondary schools in the country, meaning that we side-step one of the current issues being debated around the validity of difference-in-difference results. Our presentation of the difference-in-difference results emphasises a graphical form and a careful interpretation of any anticipation effects and a necessarily gradual implementation. The impact is necessarily gradual because, for example, it takes two years to invest two years' worth of resources in a student. We aim to be flexible in estimating the results and cautious in interpreting them. We use seven years of pupil-level data covering all state-funded school pupils in England, from 2011/12. Progress 8 was announced in 2013, and first published in 2015/16. This means we have data for four years prior to its introduction (comprising two years prior to announcement and two years after) and three years since its introduction. Importantly, the vast majority of pupils in the most recent cohort used in the analysis, who reached the end of compulsory schooling in 2017/18, would have joined the school before the announcement of Progress 8.

We judge the results to be supportive of the hypothesis that schools respond to the exact form of accountability, and in this case will realign their discretionary resources in the light of the new incentives induced by the policy change. But as has been noted, it can be difficult distinguishing between unexplained trends and a gradual causal effect with anticipation (see for example Cunningham, 2021, Angrist and Pischke, 2014). We argue that the very specific group of students that we compare makes alternative explanations less likely than the core hypothesis here. Our results are consistent with the view that schools had reacted to the previous regime of high implicit incentives for the performance of students near the threshold, and, once that incentive was removed, the borderline group made lower relative progress. The beneficiaries are students above and below the threshold, heavily weighted to the latter: our central findings show a post-reform gain of 0.01 standard deviations (SD) in GCSE English and maths for the above-borderline group and 0.06SD for the below-borderline group. Our interpretation is that when HTs were freed from the accountability-derived focus on borderline students, they chose to reallocate that "dividend" towards lower-performing students. Whether this decision was based on social norms or professional standards, or because the test scores of low-scoring students were easier to increase, is a question we cannot answer with this dataset. We return to discuss the policy implications of these findings in the Conclusion.

There is now a substantial body of evidence on school accountability (see Burgess and Greaves, 2021), briefly reviewed in the next section. The closest paper to this one in that literature is Reback (2008), who studies the distributional impact of the introduction of No Child Left Behind Act (NCLB) in the US. His paper differs from ours as the driving change there is the introduction of the new accountability and testing regime as a whole. By contrast, in this paper we consider a simple change to the metric within a very well-established system. This means we can isolate the specific impact of the reform in the system of measurement rather than the overall effect of the whole systemic change. There is also a connection to Bergbauer et al (2021) in the literature on educational testing, a conceptually distinct but practically related educational phenomenon. Their cross-country study of the uses of testing represents what might be termed a "macro" approach to this overall issue, and our paper is a complementary "micro" approach estimating the causal effect of a very specific policy change, based on a clear and explicit change with pre-reform-defined groups of students.

The remainder of the paper is organised as follows. The following section reviews the existing evidence. Section 3 presents the empirical approach including background on school accountability in England, and interpretive framework and the model we estimate. Section 4 describes our data work including how we define borderline students. Section 5 then presents the main results, estimates of treatment heterogeneity, and tests of robustness. Finally section 6 offers some conclusions for policy.

2. Evidence Review

Most of the evidence on school accountability is focussed on two questions: what are the benefits of school accountability in terms of student achievement, and what are the potential costs? The benefits arise through schools raising their game given the public performance information, and the costs come through attempts to game the metrics or even cheat.

This evidence has recently been reviewed by Burgess and Greaves (2021), and Dee (2020) offers a 30-year retrospective on school accountability in the US. Analysis of the No Child Left Behind (NCLB) accountability system in the US generates the bulk of the evidence, and Figlio and Loeb (2011) summarise this as "taken as a whole, the body of research on implemented programs suggests that school accountability improves average student performance in affected schools"; Dee(2020) concurs that school accountability in the United States did raise school performance. In England and Wales, the school accountability program was established in 1992. Burgess et al. (2013) exploit a policy that removed school performance tables in Wales but not in England, giving a better identification strategy than many contexts allow. They show that this was strongly detrimental to pupil performance in Wales.

As with any performance system, schools may seek to game the metrics embedded in the system. These take many forms: from schools focusing resources on a subset of subjects, topics, and pupils, through behavioural distortions, to outright cheating in exams. Figlio and Loeb (2011) summarize this evidence, and examples include Rouse et al. (2013), Figlio and Getzler (2006), Figlio and Winicki (2005), Bokhari and Schneider (2011), Anderson et al. (2017), Jacob and Levitt (2003) and Bertoni et al. (2013).

The current paper is focused on the impact of the particular parameters of the accountability system, and this continues to be a subject for research internationally. For example, Dee (2020) compares the likely impact of NCLB with the latest reform to the federal framework. On a closely related but conceptually different note, international comparative research using the PISA data, see Bergbauer et al (2021), shows the effects of differences in the use of universal standardised testing in 59 countries. There is also attention on mechanisms for these effects: for example, Rouse et al (2013) study how schools in Florida respond to accountability pressure.

Closest to our paper is the work of Reback (2008) who used individual pupil-level data from the 1990s in Texas to study the effects of a school accountability system, later transferred to the Federal level as NCLB. His results suggest that "schools respond to the accountability system by taking actions which influence the distribution of student achievement." Specifically, he shows that schools will target resources on students whose scores matter disproportionately for the overall accountability-relevant performance of the school.

3. Empirical Approach

a) School Accountability in England

A school accountability system typically involves regularly published performance indicators, which provide information for parents and to school authorities. It also provides implicit incentives to schools because this publication can exert pressure on the leadership of low-performing schools. In England, school performance tables have been published annually since 1992 following the Education (Schools) Act of 1992. Whilst there have been numerous changes to the content and presentation of the tables[2], they have been an ever-present part of the school system, and the core purpose and content has remained the same.

This paper focusses on the most important change yet to the content, the introduction of Progress 8 as a headline measure of school performance in 2015/16, which we describe next.

[2] We document these in an earlier report (Burgess and Thomson, 2020)

Over almost all of this time, the main headline measure has involved a threshold – the fraction of a school's pupils hitting some exam benchmark – achieving at least 5 subjects ("exams") marked at least grade C. The threshold feature, unsurprisingly, implies strong incentives for the school to help the pupils just on the boundary of getting a C or a D grade. This has been described as distorting schools' behaviour, forcing schools to focus on pupils around this borderline. But this is a feature, not a bug: a built-in design feature which, by focussing incentives, can be very powerful (see Burgess, 2013). If there is a level of achievement that wider society feels is extremely important for everyone to reach, then it makes sense to set up a scheme that offers very strong incentives to do that – that focusses the incentive around that minimum level. This is precisely what a threshold scheme does.

This was the headline performance metric in place before the Progress 8 reform. After the reform, another threshold measure, the percentage of pupils achieving grade C/4 or higher in GCSE English and maths, was published although this was given much less emphasis than Progress 8.

b) Introduction of Progress 8

The Department for Education (DfE) announced in October 2013 that a new set of 'headline' measures would be published in January 2017, summarising school performance in the 2015/16 academic year (Department for Education, 2013): Attainment 8 (A8), and Progress 8 (P8).

An individual pupil's A8 score is derived by allocating points to grades achieved in certain qualifications. There were eight subject 'slots' in total: one for English, one for maths, three for the 'Ebacc' subjects[3] (sciences, humanities[4], modern and ancient languages) and three 'open' slots for any other eligible qualifications. English and maths were double-weighted so there were effectively 10 slots in total.

A pupil's P8 score is a progress version of A8. The calculation is a simple one: banding pupils based on their prior attainment (see Burgess and Thomson, 2013), computing the average A8 score for each band, and thereby deriving the 'expected' A8 score for each pupil based on their prior attainment. Finally, the gap between this and her actual A8 score is the P8 for an individual pupil. The school level P8 metric is simply the average of this over all pupils in the school; this was introduced as the new lead accountability measure.

This new measure therefore entailed two main changes. First, it is a value-added type measure, taking into account each pupil's prior attainment when they joined the school. Second, and the focus of this

[3] https://www.gov.uk/government/publications/english-baccalaureate-ebacc/english-baccalaureate-ebacc
[4] Geography, history or ancient history.

paper, it is calculated as a simple average of all pupils in the school, it is not a threshold measure. Schools can affect their performance outcomes in two ways – by changing how they teach, or changing who they teach; that is, by reallocating their teaching resources across their pupils, or by adjusting their admissions rules. Here we focus on the former, and we sidestep the latter by using data only from pupils assigned to schools during the prior regime[5].

c) Research questions and hypotheses

Our overall research aim is to advance our understanding of how school accountability systems influence school behaviour and therefore pupil outcomes. Specifically, we evaluate the impact of the systemic change of the introduction of P8 on pupil outcomes.

To interpret our results, we sketch out a modelling framework. Every Headteacher (HT) has some discretionary resources that can be assigned to different groups of pupils; this includes how to assign the most effective teachers, smaller class sizes, or additional learning resources. Consider the HT's decision on how to allocate these over pupils across the range of abilities in the school. The two key factors are the relative weight that the HT places on the outcomes for different ability students, and the relative cost of raising attainment for a particular ability group. On the first, this of course might be entirely even – the school values the progress of all its pupils equally. But there are other possibilities. For example, it may be that the HT particularly wants to raise attainment of low ability pupils for 'social mission' reasons. Alternatively, for example, being aware of strong pressure from parents of high ability pupils, the HT may favour them by assigning the most effective teachers to the top sets. Equally, the most effective teachers, whom the HT needs to keep happy, may want to work with the top sets, with the same outcome. In any case, we can imagine a weight for each ability group reflecting the HT's preferences. On the cost side, it may be very simple: that the cost in extra resources of raising attainment may be the same for all ability groups. Or it may be that higher ability pupils are easier to raise up based on a greater initial skill set; or that higher ability pupils are more costly to raise because of ceiling effects. A comparison of these two functions across ability groups will highlight the ideal ability groups for HTs to target their most effective teaching resources on.

The P8 reform removes the additional strong incentives for gains for pupils close to the threshold. In the absence of any new policy-driven focus group of pupils the comparison of costs and benefits is the best way of allocating resources. Because neither of these two functions are well evidenced, we remain agnostic as to what to expect. What is incontrovertible, however, is that HTs were strongly

[5] There are a small number of areas in England that operate a "middle school" system, when pupils join the schools in which they will take their GCSEs at age 13/14, rather than 11/12. For later cohorts this would have been after the P8 announcement.

aware of the allocation incentives from the prior regime, and would therefore have been equally aware of their removal and the resulting 'clean slate' in terms of what to do instead.

One additional factor is that while the C/D threshold was no longer a direct part of schools' incentives, it remained a key benchmark for pupils and continued to be published in School Performance Tables. Achieving at least a grade C (later called grade 4) passes in English and mathematics was essential for progression to the next stage of education, and/or in the labour market; this might have meant that for some schools there was a residual indirect emphasis on the borderline group.

d) Empirical Model

The introduction of P8 affected all relevant schools all at once, there was no variation in timing[6] and no reversion to the old system. However, there are strong *a priori* grounds for believing that different groups of pupils will be differentially affected: pupils around the prior regime's performance threshold, as opposed to those above or below it. This suggests a difference-in-difference approach, modelling the effect of the policy change interacted with pupils in different groups. In our analysis these groups are pre-determined for the reform: characterized by pre-reform definitions and estimated using only pupil characteristics that were fixed before the reform. The key group are those considered to be borderline for achieving the key accountability metric pre-reform, pupils who were marginal to the sharp threshold of five A*- C grades. We allow for schools to differentially shift resources to pupils above the threshold and to those below the threshold.

We estimate the following as our main specification for pupil *i* in school *s* in academic-year *t*:

$$g_{ist} = \beta.X_i + \mu_{st} + \delta.\tau + (\sigma_0 + \sigma_1.\tau) * b_{is} + (\alpha_0 + \alpha_1.\tau) * a_{is} + \varepsilon_{ist} \quad (1)$$

The dependent variable, g_{ist} is the test score outcome for a pupil (see the Data section below). We define group dummies b_{is} and a_{is}, where a_{is} is equal to 1 if pupil *i* in school *s* is denoted "above the borderline"; and b_{is} is equal to 1 if pupil *i* in school *s* is denoted "below the borderline"; τ is the "after" dummy with τ =1 in the post-reform years (2016 to 2018), and zero otherwise. The key difference-in-difference terms are τ *x* b_{is} and τ *x* a_{is}, with coefficients σ_1 and α_1 We supplement these in our data with pupil characteristics[7], X_i, and μ_{st}, a set of fixed school-by-year effects. We standardise prior attainment (Key Stage 2) scores for each year and fit them as third-degree polynomials.

[6] A small number of schools chose to opt in to P8 a year earlier than the official date. We check the robustness of our results to this group by running a specification dropping them out.

[7] Note we do not place a lot of weight on the interpretation of the coefficients on these variables. They are only separately identified from the above and below variables (a_{is} and b_{is}) by functional form. We also provide a specification without these controls.

The identification of causal effects using the difference-in-difference model set out in equation 1 assumes parallel trends in outcomes between the groups of pupils prior to the policy change. We examine the pre-reform trends in greater detail in Section 5.a. The model can be extended to allow for group-specific trends (see for example Angrist and Pischke, 2014) in our multi-year context:

$$g_{ist} = \beta.X_i + \mu_s + \delta.\tau + (\sigma_0 + \sigma_1.\tau + \varphi.* t) * b_{is} + (\alpha_0 + \alpha_1.\tau + \theta * t) * a_{is} + q_t \quad (2)$$

Here we have added group specific linear time-trends, $a_{is} * t$ and $b_{is} * t$, and a set of common year effects, q_t.

e) Empirical implications of anticipation and a gradual transition between regimes

The transition period began in October 2013 with the announcement of the policy, with the new performance measure to be first applied for outcomes in the year 2015/16. How might schools react?

In October 2013, it seems likely that at least some of the big prioritisation decisions for the year 2013/14 would have been taken (for example the assignment of teachers to classes), and so we would not really expect any impact on attainment for the 2013/14 Year 11 cohort. It seems more likely that schools would be able to change policies (if they wished to) from 2014/15. This might only affect the schools most attuned to the incentive structure and keenest to change. This would proceed as follows; the P8 reform could affect exam outcomes:

School year:	Potential effect of P8 reform on schools' decisions:	Number of years of potential additional investment in focus students:
2014/15	1 year of change in schools' prioritisation decisions, presumably for year 11 students	1
2015/16	2 years of change in schools' prioritisation decisions, presumably for years 10 and 11 students	2
2016/17	3 years of change in schools' prioritisation decisions, presumably for years 9, 10 and 11 students	3
2017/18	4 years of change in schools' prioritisation decisions, presumably for years 8, 9, 10 and 11 students	4

All these students would have been assigned to schools under the old accountability regime[8]. Pupils taking GCSEs in the summer of 2016 would typically have joined their school in September 2011, and chosen their school in 2010, and pupils taking GCSEs one year later joined in 2012 and chosen in 2011. The period we could describe as fully post-policy-change would be from when all secondary school years were under the new regime, which would start with the 2018/19 GCSEs. However, those pupils joining their school in Year 7 (the usual time at which admission to secondary school occurs) would have done so in September 2014, after the announcement of Progress 8.

What are the implications of this for the empirical approach? Under the hypothesis studied here that schools react in an optimising way to the accountability framework they work under, we would expect to see a gradual build-up of change as schools switch to the new investment strategy and pupils have more and more years under the new approach. It is important to be clear that such a gradual change is simply due to the passage of time, rather than any slow or reluctant reaction by schools; it takes two years for pupils to have two years of priority investment. We expect to see zero change in exam outcomes in 2013/14 exams, through a small effect in 14/15, bigger in 15/16, and so on through 18/19, until outcomes stabilise.

Note that this is a very different expected profile to the archetypical difference-in-differences model in which the policy change produces an instant and on-going effect (see for example the many figures illustrating this in Angrist and Pischke (2014) and Cunningham (2021).

This has important implications for our evaluation of the fit of our model, principally in terms of the standard analysis of prior trends and placebo tests. Essentially, the issue is this: the optimising model of schools reacting to new incentives implies (as above) a necessarily gradual reaction to the policy, some portion of which is quite likely to happen shortly before the formal implementation start date. In this case, the data will present as differential prior trends and pre-implementation effects. But these patterns are precisely those that are taken to cast doubt on the validity of a difference-in-differences analysis. One solution would be to count the policy change date as the announcement of the policy, October 2013, but this runs up against the data problem that that moment is only one year after the

[8] There are three things worth noting. First, some pupils will have changed school before the end of Year 11 and some of these moves will have happened after the policy was announced; second, schools can still game the system by excluding (formally or informally) students before the end of Year 11; and third, there was a small increase in special types of schools admitting at age 14 (e.g. UTCs and studio schools) during our observation window.

adoption of “comparable outcomes” policy that ended grade inflation and initiated a period of stable score distributions, potentially prejudicing the before/after comparison.

4. Data

a) National Pupil Database

We use pupil-level administrative data from the National Pupil Database (NPD), maintained by the Department for Education. The data contains pupils’ results in GCSEs and other approved qualifications at the end of compulsory schooling (Key Stage 4), usually at the age of 16. These records have been matched to details of prior attainment in tests and teacher assessments at the end of primary school (Key Stage 2), usually at age 11. Data on pupils’ characteristics (gender, free school meal eligibility, ethnicity, month of birth and whether their first language was English) is matched in from the School Census. Data on local neighbourhoods, such as the Income Deprivation Affecting Children Index (IDACI) was also added to pupil records as the lower super output area (LSOA) in which they reside are contained in NPD. Our dataset uses data on all pupils who reach the end of Key Stage 4 in state-funded mainstream schools between 2011/12 and 2017/18.

This data is supplemented with additional data about schools, including governance and admissions policy (*Get Information About Schools*, UK Government, 2020).

b) Measuring pupil attainment

Producing a set of pupil (and therefore school) performance indicators on a consistent basis across our analysis period is not straightforward, and needs to take account of changes in: (i) the mapping of grades to scores, (ii) the set of qualifications and their “equivalences” that count for the performance tables, and (iii) the nature of the performance tables methodology. Details of what we did are in Appendix 1. For pupil outcomes we use three indicators for which the definitions and measurement have been relatively stable over this period:

- Average points score in English and maths (English and maths APS).
- The achievement of five or more A*-C grades (or equivalent) including English and maths (5ACEM); this is the headline accountability measure under the old regime.
- Mean grade in GCSEs (Mean GCSE)

English and maths APS is our primary outcome, because (almost) all pupils enter English and maths, and because entries in English and maths are unaffected by other reforms which took place during our observation window. Grades in GCSE English language (or combined English language and

literature[9]) are converted into points[10] and then averaged. GCSEs were graded A*-G since their inception in 1988 until 2015/16. The appearance of reformed GCSEs in Key Stage 4 data for 2017 causes us a headache for our APS measure. However, we use a simple transformation to map the new 9-1 grades onto the previous points scale (see Appendix 1). The new grades were designed such that grades 3-1 correspond to the former D-G range, 6-4 corresponds to the former B-C range and 9-7 corresponds to the former A*-A range. English and maths APS sits somewhere between low stakes and high stakes. Although not published, scores in English and maths compose 40% of the Attainment 8 measure. Furthermore, almost all pupils enter English and maths and this has been the case for the full period of our analysis.

We also use a lower stakes indicator that is less affected by the importance attached to English and maths, mean GCSE grade. Although it has been published since 2010/11, it tends to have lower prominence in comparison to other indicators.

The means and standard deviations of the three key outcome measures and the key pupil-level controls for all pupils at the end of Key Stage 4 in state-funded mainstream schools are shown in Table 1.

c) Defining 'borderline' pupils

We cannot know which pupils a school thought of as being borderline. Such a judgement likely included inputs from internal low-stakes tests, teacher assessments and so on, and might evolve over time. We produce a simple proxy, which we assume is correlated with schools' own views, based on a pupil's prior attainment and two key time-invariant factors, gender and month of birth, that predict GCSE scores well. We estimate for every pupil in our dataset their probability of achieving five or more A*-C grades including English and maths (5ACEM), the key attainment threshold pre-reform, and define a range of the distribution of fitted probabilities as distinguishing borderline pupils.

We take an *ex post* approach and estimate the probability of 5ACEM retrospectively, that is, using the actual GCSE scores for each pupil. Taking each year in the dataset in turn, and looking backwards from their realised GCSE score, we use logistic regression to estimate the relationship between these scores

[9] This was a single GCSE available until 2016. Since 2017, only separate GCSEs in English language and literature have been available.

[10] Grade A*=58; A=52; B=46; C=40; D=34; E=28; F=22; G=16, U=0. Pupils not entered are assigned 0 points. Equivalent scores for pupils entering AS-levels are also used. A pupil's highest score is used.

and the pupil factors[11]. We do this in preference to an *ex ante* approach in order to smooth out fluctuations over time in the relationship between the outcome and the predictors.

The final step is to specify what part of the distribution of probabilities counts as 'borderline'. In our main specification we choose the range 40% to 60%, and refer to pupils with a higher probability as being in the 'above' group and those with a lower probability as the 'below' group. In our tests of robustness, we also show the effect of widening (and narrowing) this window.

The outcome of this modelling is displayed in Figure 1 for the full length of available data: the percentage of pupils in each of these three groups in each year[12]. The percentage of borderline pupils is relatively stable over the period 2012 to 2018, generally forming 13% to 14% of each cohort. The exception is 2015. This cohort was affected by a boycott of Key Stage 2 tests in 2010. For around a quarter of pupils, teacher assessment data has been used in place of test data to assign pupils to groups and this has introduced a slight degree of turbulence into the series.

If we cut the data by school, we can see that over our analysis window, 2012-2018, the variation between schools in their fractions of borderline students is not large, presented in Table 2. The 10th percentile of fraction borderline students is generally 8%-9%, and the 90th percentile is 18%-19%. The 2015 cohort was affected by the 2010 Key Stage 2 boycott and so appears to be an outlier. Very few schools have hardly any borderline students and in very few schools do they account for more than a fifth.

d) Analysis period

Our analysis period is GCSE exam results taken in 2012 (so at the end of the school year 2011/12, based on pupils' learning from 2010 through 2012), each year through the exams at the end of the school year 2017/18. Progress 8 was announced in October 2013, and first published in January 2017 relating to the exams taken in June 2016.

We start from 2011/12 because this was the first year after a very impactful change in assessment practices. This was the first year that a policy called 'comparable outcomes' was officially applied to English and maths GCSEs (Ofqual, 2011), and signalled the end of grade inflation. Relatedly, the percentage of pupils in the below-borderline group largely levelled out as indicated in Figure 1. These two factors mean that the core assumption of parallel trends is more likely to be satisfied during this

[11] We also ran the *ex-ante* version, see (Burgess and Thomson, 2020), which we do not report here because it produced almost identical results.

[12] We do not include pupils without KS2 results (for instance those arriving from overseas during their secondary education) in our analysis.

period; we examine this below. The final cohort of pupils admitted to Year 7 in secondary schools prior to the announcement of Progress 8 entered GCSEs in 2017/18.

We standardise our attainment measures to standard deviation units over the analysis period 2012 to 2018, so the results are presented in effect size units.

5. Results

a) Main Results

We present our difference-in-differences results in Table 3, summarising the graphical evidence presented in the following sub-section. This uses average points score in English and Maths as the dependent variable and our base definition of 'borderline' as pupils with a 40-60% probability of achieving the threshold, and includes different specification of controls. Standard errors are clustered at school-by-year level. For each specification, we report coefficients on the below group, the above group, and the post-reform dummy. Note that the simple group dummies cannot be straightforwardly interpreted: they are strongly correlated with the pupil characteristics also included in the regression and are only separately identified by functional form (the logistic regression determining the borderline group).

The focus of interest is the difference-in-differences coefficients. Column (1) reports the base model with no additional controls. This shows a post-reform increase in test scores relative to borderline pupils of 0.08 of a standard deviation (SD) for the below-borderline group, and essentially no effect for the 'above' group.

The second column adds school-by-year dummies, thereby controlling for aggregate time effects, time-invariant school effects, and school-year specific effects in a very flexible way. The difference-in-difference estimates barely change. The third column adds pupil characteristics, listed below the table, but removes the school-by-year dummies. This has two effects. First, as expected, this makes a big difference to the simple estimated group effects, as they are simply non-linear functions of some of the characteristics. Second, and more importantly, the difference-in-difference coefficient for the 'above' group now becomes positive and statistically significant; the coefficient for the 'below' group declines slightly.

The fourth column presents our full specification with both pupil characteristics and school-year effects, we find that the post-reform effect for the 'above' group is 0.012 SD and for the 'below' group is 0.072 SD. The average of these two terms weighted by the number of pupils in the above and below

groups, 0.032SD, is the additional value that the borderline group experienced prior to the introduction of Progress 8.

The fifth column presents the same specification but this time the standard errors are two-way clustered by year and school. This increases the size of the standard errors, with the post-reform effect for the above group becoming non-significant.

In Table 4 we present results for the two other outcomes, the achievement of 5 or more A*-C grades (or equivalent) including English and maths (AC5EM) and mean GCSE grade in all subjects. These also show positive, albeit smaller, effects for the 'below' group of 0.041 SD and 0.048 SD respectively. For the 'above' group the effects are close to zero at -0.01SD and 0.005SD respectively.

We discuss the quantitative effect of these results below, but three immediate conclusions are that: (i) the use of the threshold measure made a statistically significant difference to school outcomes, we assume arising from changed school behaviour, focusing their resources on the incentivised group of pupils, and (ii) when that incentive was reduced, schools reacted, and (iii) redistributed resources to the non-borderline groups more heavily weighted towards the below-borderline group.

b) Prior trends

The two key assumptions for a difference-in-differences approach to yield a valid causal estimate are that there is no movement between groups, and that the different groups considered would have had common outcomes trends after the policy change. By definition, there can be no movement of a pupil between groups after the reform, as that derives from our non-time-varying estimation and is defined by pre-reform covariates. We now address the issue of common trends.

Figure 2 **Error! Reference source not found.**shows the results of estimating:

$$g_{ist} = \beta . X_i + \mu_s + \lambda_t + \sum_{\tau=2012}^{2018} \delta_\tau . B_{i\tau} \quad (3)$$

where we present the coefficient (δ) for the borderline group (B) interacted with each year in turn 2011/12 through 2017/18 (with 2012/13 acting as base year), along with the associated 95% confidence intervals, clustered by school and year.

The pattern from 2014/15 onwards fits with the hypothesis set out above, a gradual decline each year after 2013/14 for the borderline group. This is consistently downward, but particularly marked in 2017/18. It is also worth noting that there is some instability in our outcome measure prior to 2014/15 as a result of the rise and fall in multiple entry (resits) in English and maths (Appendix B). The result for 2018 (-0.06 SD) is equivalent to around 9% of a grade at GCSE.

In Figure 3 we separate out the above-borderline group and the below-borderline group, allowing for different effects. We see that both groups tend to gain from 2013/4, the above group marginally, the below group more dramatically so. This is consistent with the hypothesis that changing prioritisation policies of schools will have more effect each year from 2013/4 onwards as pupils are "treated" for successively more of their school careers. Again to emphasise, the gradual impact can simply arise from the passage of time: pupil outcomes diverging from differential investment over 1, 2, 3, … years.

These patterns are consistent with the hypothesis that the reform to the accountability system changed schools' incentives for targeting interventions and that this in turn led to changes in pupil outcomes. We are clear that these patterns could also fit other, non-causal, stories, or that there simply are unexplained trends starting roughly around the time of the reform we are focussing on and which are just unluckily coincident. As has been noted (Cunningham 2021, Angrist and Pischke, 2014), distinguishing between unexplained trends and a gradual causal effect with anticipation can be difficult. Working in favour of the hypothesis we set out is that the borderline group we define is quite narrow and "specific" – that is, it is relevant only in relation to the accountability process for schools.

c) Heterogeneous Treatment Effects

Schools' responses to the changed incentives are likely to depend on their context. For example, highly selective (Grammar) schools have very few (if any) pupils at the C/D border so the prior regime would have been irrelevant for them. We expect that the impact of the reform would be greater in cases where schools had reacted more to the old regime. We consider school responses by:

- different historical levels of pupil attainment and performance.
- different degrees of local competition
- different portfolios of types of qualifications entered. Progress 8 encourages schools to enter pupils for particular qualifications (e.g. GCSEs in Ebacc subjects) which tend to be graded more severely than alternative non-GCSE qualifications.
- different proportions of borderline pupils.

Note the first two of these might be thought exogenous, but the latter two are clearly chosen by the schools. For those results, the interpretation is different as we cannot claim they are necessarily causal.

Results are summarised in Figure 4. First, we consider schools under strong pressure from being near the 'floor standard' that existed before the P8 reform[13]. For them, the desire to increase the performance of borderline pupils prior to the reforms was likely to be intense. We define this group as schools having performance in the previous year between 35% and 45% 5ACEM. Indeed, the largest effects can be observed in schools that were close to the floor standard. Once the pressure to focus on borderline pupils was removed, the attainment of the above-borderline and below-borderline groups improved more so than in other schools.

Second, we split by school performance as approximated by a measure of contextual value added[14], and interacting the lowest quintile and the highest. The effects of the reform were smaller in schools with high contextual value added, suggestive that there was less of a focus on borderline pupils prior to the introduction of Progress 8 in these schools.

Third, we use a metric that has been taken to characterise strategic behaviour by schools, namely the extent to which they use (supposedly much easier) non-GCSE qualifications. We find that schools making greater use of non-GCSEs also reacted more strongly to the reform.

Fourth, we consider competitive pressure on schools from the density of alternatives available to parents, measured here by the number of other state-funded mainstream schools within a 3km straight-line distance of the focus school. We know from Burgess et al (2013) that the presence of school performance tables causes schools to focus on and improve their measured performance. Schools for which these competitive forces felt more immediate might be expected to maximise their chances in the market by strongly engaging in prioritising the borderline students. We would therefore expect the removal of the threshold effect to produce bigger changes away from the borderline group in highly competitive areas. Although we see this for the above-borderline group, we do not for the below-borderline group.

Fifthly, we look at variations between schools in the fraction of borderline pupils. We emphasise again that this is an endogenous variable, both school performance and admissions will affect this. This might matter for the following reason: schools with just a few borderline pupils would be well placed to channel resources as they could target that quite intensively on the few borderline pupils. A school in which a substantial fraction are borderline however, would it find it much less cost-effective. In fact, we do not see any material differences for below-borderline pupils with respect to the fraction of borderline pupils at a school. This is consistent with column (7) of Table 4. However, there is a slightly

[13] At least 40% of pupils achieving five or more A*-C grades at GCSE (or equivalent) including English and maths (5ACEM).

[14] The residual from regressing the outcome measure on prior attainment plus the full set of pupil characteristics.

larger effect for the above-borderline group. This would be consistent with our expectation as these schools would have previously had the most to gain in terms of published performance indicators by focusing on the borderline group.

d) Robustness

We review the impact of some of our decisions on data and modelling on the results, shown in Table 5. Columns (1) and (2) show the effect of varying the number of pre-reform years. In column (1) we add data for 2010/11, the year prior to the application of comparable outcomes in GCSE English and maths. The difference-in-difference estimates barely change for the below group but there is some slight change for the above group.

Column (3) excludes the 324 schools which opted early into Progress 8 in 2014/15, which also makes little substantive difference. Columns (4) and (5) adjust the definition of borderline pupils: respectively using a broader definition of the 'borderline' group (pupils with a chance of hitting the threshold between 30% and 70%), and a narrower one (45% and 55%). The first of these makes very little material difference. Widening the borderline group increases the effect of the reform by 0.011 SD for the below group and narrowing it reduces it by 0.008 SD. To reiterate, we can only make assumptions about which group of pupils the school thought of as borderline.

Finally, in column (6) we show the effect of fitting a linear trend in outcomes for the above and below and group (equation 2 in Section 3.6). This increases the effect for the above group by 0.012SD and reduces the effect of the below group by 0.05 SD.

More broadly than these specific data decisions, one issue is the degree to which we can distinguish the effect of the P8 reform from the effect of a number of changes to schools' environment. The dates do not coincide directly with the reform we study here for any of these other changes, but there is some degree of overlap in the longer periods around them. The point is that none of these would be expected to affect just the borderline group as defined here, and so we are reasonably confident that we have identified the effects of the specific P8 reform.

Understanding the Mechanism Whilst these econometric results capture the causal statistical relationship between the P8 reform and pupil achievement, the details of how the change in resource allocation affected achievement are of interest. In Burgess and Thomson (2020) we report on a survey we ran of over 400 school leaders and teachers in England to find out more about how they responded to the introduction of Progress 8. Summarising, schools' responses were varied, but the results suggest

a general shift away from running intervention sessions aimed specifically at borderline pupils towards a greater emphasis on pupils anywhere in the ability distribution who were judged to be falling behind.

6. Conclusion

School accountability is now a widespread policy tool and fine-tuning the parameters of the system for different policy outcomes is an important design issue. In fact, in its World Development Report for 2018, the World Bank (2018) urges greater use of student testing "[t]here is too little measurement of learning, not too much" (p. 17). We contribute to this evidence by using seven years of attainment data on secondary schools in England to explore schools' reactions to significant changes to their accountability framework. The results are consistent with the view that schools had reacted to the previous regime of high implicit incentives for the test scores of a particular group of students. Once that incentive was removed, that specific group appear to make less relative progress, and other groups made faster progress. The effects are not trivial: our headline findings show a post-reform gain of 0.01SD for the above-borderline group and 0.07SD for the below-borderline group.

We have been cautious in presenting these results noting the issue of trends subsequent to announcement but before implementation. We judge the results to be supportive of the hypothesis that schools responded strongly to the changes in the accountability metric. The results are robust to a variety of other specification tests.

These results have a bearing on the test score gap between disadvantaged pupils[15] and their peers. Our findings show a post-reform improvement of around 0.01 SD for disadvantaged pupils, which can be decomposed in terms of the accountability-relevant groups[16].

This analysis has messages for policy. First, our results suggest that the introduction of Progress 8 had the intended effect of shifting schools' focus away from students who were marginal to the previous accountability threshold. The effect is not trivial but nor is it a dramatic change. In that sense, the policy "worked".

Second, this reinforces the view that accountability measures are an effective policy tool. They do not impinge directly on schools' operational autonomy, unlike explicit Ministerial directives, but they do adjust the incentive structure that schools face. Our results show that this can be effective in changing behaviour. The setting, and occasional re-setting, of the accountability framework seems an appropriate role for Government – it is the practical expression of its view of what society deems

[15] Defined here as pupils eligible for the Pupil Premium (PP), based on family income.
[16] Details in Appendix 2.

valuable in education, of what schools 'ought' to do. Problems clearly arise if the framework is changed very frequently so that schools do not have a stable environment for planning.

Third, problems can also arise if different parts of schools' incentives pull in different directions. The previous accountability regime was based on the 5ACEM threshold, so schools were strongly incentivised to maximise the fraction of their pupils that achieved this. This drive meshed well with the goal of the typical pupil because for her passing that threshold was key to access to higher or further education and to the job market. Schools could allocate their resources knowing that the goal of doing well by their pupils and the goal of doing well on the performance metrics were closely aligned. In the new regime, that is less true. Access to higher education and to jobs is still to an extent dominated by the 5ACEM threshold, and this may mean that schools are partially conflicted, and that a goal for the school of keeping the 5ACEM "pass rate" high is still important to them. It may be that the labour market and HE admissions will respond and place more emphasis on P8 scores, or it may be that these two goals for schools will remain in tension.

It is not the case that schools in general simply "try to do what's best" for their pupils; rather, they respond to the details of the incentive structure they are given. Whether this is seen as positive or negative depends on the nature of the incentives given, and on the social value placed on the educational achievements of the pupils favoured by the incentive.

References

Anderson, P. M., Butcher, K. F., & Schanzenbach, D. W. (2017). Adequate (or adipose?) yearly progress: Assessing the effect of "No Child Left Behind" on children's obesity. *Education Finance and Policy*, *12*(1), 54–76.

Angrist, J. D. and Pischke, J. S. (2014) *Mastering 'metrics: The path from cause to effect*. doi: 10.5860/choice.189854.

Bergbauer, A.B., E. A. Hanushek, and L. Woessmann (2021). 'Testing'. *Journal of Human Resources (forthcoming)*. https://drive.google.com/file/d/12pNJcbqTX6UXUiu5JI6Dnuoor39PBQ2c/view (Accessed: 19th August 2022).

Bertoni, M., Brunello, G., & Rocco, L. (2013). When the cat is near, the mice won't play: The effect of external examiners in Italian schools. *Journal of Public Economics*, *104*, 65–77.

Bokhari, F. A. S., & Schneider, H. (2011). School accountability laws and the consumption of psychostimulants *Journal of Health Economics*, *30*(2), 355–372.

Burgess, S (2013) Threshold measures in school accountability: asking the right question. CMPO Viewpoint Blog. https://cmpo.wordpress.com/2013/09/25/threshold-measures-in-school-accountability-asking-the-right-question/ (Accessed: 19th August 2022).

Burgess, S., & Greaves, E. (2021). School Choice and Accountability. *Oxford Research Encyclopaedia of Economics and finance*. https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-650 (Accessed: 19th August 2022).

Burgess, S. and Thomson, D. (2013) *Key Stage 4 Accountability: Progress Measure and Intervention Trigger*. Available at: http://www.bristol.ac.uk/media-library/sites/cubec/migrated/documents/report11.pdf . (Accessed: 19th August 2022).

Burgess, S. and Thomson, D. (2019) 'The impact of the Wolf reforms on education outcomes for lower-attaining pupils', *British Educational Research Journal*. doi: 10.1002/berj.3515.

Burgess, S. and Thomson, D. (2020) School accountability and fairness: Does 'Progress 8' encourage schools to work more equitably? Nuffield Foundation Report https://www.nuffieldfoundation.org/wp-content/uploads/2020/12/Full-report-School-accountability-and-fairness.pdf

Burgess, S., Wilson, D. and Worth, J. (2013) 'A natural experiment in school accountability: The impact of school performance information on pupil progress', *Journal of Public Economics*. doi: 10.1016/j.jpubeco.2013.06.005.

Cunningham, S. (2021). *Causal Inference: The Mixtape.* Yale University Press. doi: 10.2307/j.ctv1c29t27

Dee, T. (2020). *Learning from the Past: School Accountability before ESSA. A Background Paper for the Hoover Education Success Initiative.* https://www.hoover.org/research/learning-past-school-accountability-essa (Accessed: 19th August 2022).

Department for Education (2013) *Reforming the accountability system for secondary schools Government response to the February to May 2013 consultation on secondary school accountability*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/249893/Consultation_response_Secondary_School_Accountability_Consultation_14-Oct-13_v3.pdf (Accessed: 19th August 2022).

Figlio, D. N., & Getzler, L. S. (2006). *Accountability, ability and disability: Gaming the system. *Advances in Applied Microeconomics*, *14*, 35–49.

Figlio, D. & Loeb, S. (2011). School Accountability in Hanushek, E., Machin, S. & Woessmann, L. (eds.) *Handbook of the Economics of Education,* edition 1 volume 3 pp. 383-421. Elseveier. doi: 10.1016/B978-0-444-63459-7.00005-**1**

Figlio, D. N., & Winicki, J. (2005). Food for thought: The effects of school accountability plans on school nutrition. *Journal of public Economics*, *89*(2–3), 381–394.

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

Ofqual (2011) *GCSEs and A Levels in Summer 2012: Our approach to setting and maintaining standards*. Available at https://dera.ioe.ac.uk/15397/1/2012-05-09-maintaining-standards-in-summer-2012.pdf (Accessed: 19th August 2022)

Reback, R. (2008) 'Teaching to the rating: School accountability and the distribution of student achievement', *Journal of Public Economics*.

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 (2): 251-81.

UK Government (2020) *Get information about schools*, *Webpage*. Available at: https://get-information-schools.service.gov.uk/ (Accessed: 19th August 2022).

World Bank. 2018. World Development Report 2018: Learning to realize education's promise. Washington, DC: World Bank.

Woessmann, L. (2016). 'The Importance of School Systems: Evidence from International Differences in Student Achievement'. *Journal of Economic Perspectives* vol. 30(3) pp. 3–32.

Appendices

Appendix 1: Equating legacy and reformed GCSE grades

The qualifications pupils are observed to have entered, and the performance indicators calculated for them in NPD, are dependent on the prevailing accountability framework of the day. For example, prior to the introduction of Progress 8, GCSE grades were "scored" using a scale that ranged from 16 points for grade G to 58 points for grade A* with the intervening grades scored at 6 point intervals. From 2016, they were scored 1 point for grade G to 8 points for grade A*. This was in fact the original scoring system used until 2003. It was a brief respite since the scores changed again in 2017 to accommodate reformed GCSEs which were graded on a different scale (9-1).

In addition, the response to the Wolf Review led to changes in equivalence for some qualifications and others no longer being counted at all in school performance tables from 2014 onwards (Burgess and Thomson, 2019).

We therefore transform pupil (and therefore school) performance indicators onto a consistent basis. There are two main aspects to this. Firstly, we calculate indicators for 2014 to 2018 using the 2013 Performance tables methodology. This means we include all approved qualifications, not just those that were deemed eligible following the Wolf Review, and apply the points scoring system that prevailed in 2013 to results from 2016, 2017 and 2018. There is a necessary caveat here: schools would have responded to the prevailing accountability incentives to have entered particular qualifications. We cannot readily adjust for these different decisions, however we can attempt to use outcome measures that we believe are relatively stable in our analysis.

Reformed GCSEs (graded 9-1) were first awarded in English and maths in 2017. In the absence of any existing conversion of these grades into the points scale used by the Department for Education between 2005/06 and 2015/16, we assign points to 9-1 grades as follows:

13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
			G						F						E						D						C						B						A						A*			
						1									2						3				4			5				6					7				8					9		

This yields a reasonably similar means and standard deviations of points for 2017 compared to 2016 for all three parts of the distribution (Table A1).

Table A1: Average point scores for English and maths, 2016 and 2017

	Average points score		Standard Deviation	
	2016	2017	2016	2017

English	D-G	31.0	31.0	4.7	4.7
	B-C	42.6	43.0	3.0	3.2
	A*-A	53.5	53.6	3.1	3.3
Maths	D-G	28.8	28.8	6.4	5.8
	B-C	42.4	42.2	2.9	3.1
	A*-A	54.4	54.1	4.5	4.3

Appendix 2: Implications for the disadvantage gap

What does this reform imply for the distribution of pupil achievement, and in particular for the achievement gap between disadvantaged pupils and their better-off peers? We define the former as pupils eligible for the Pupil Premium (PP), based on family income.

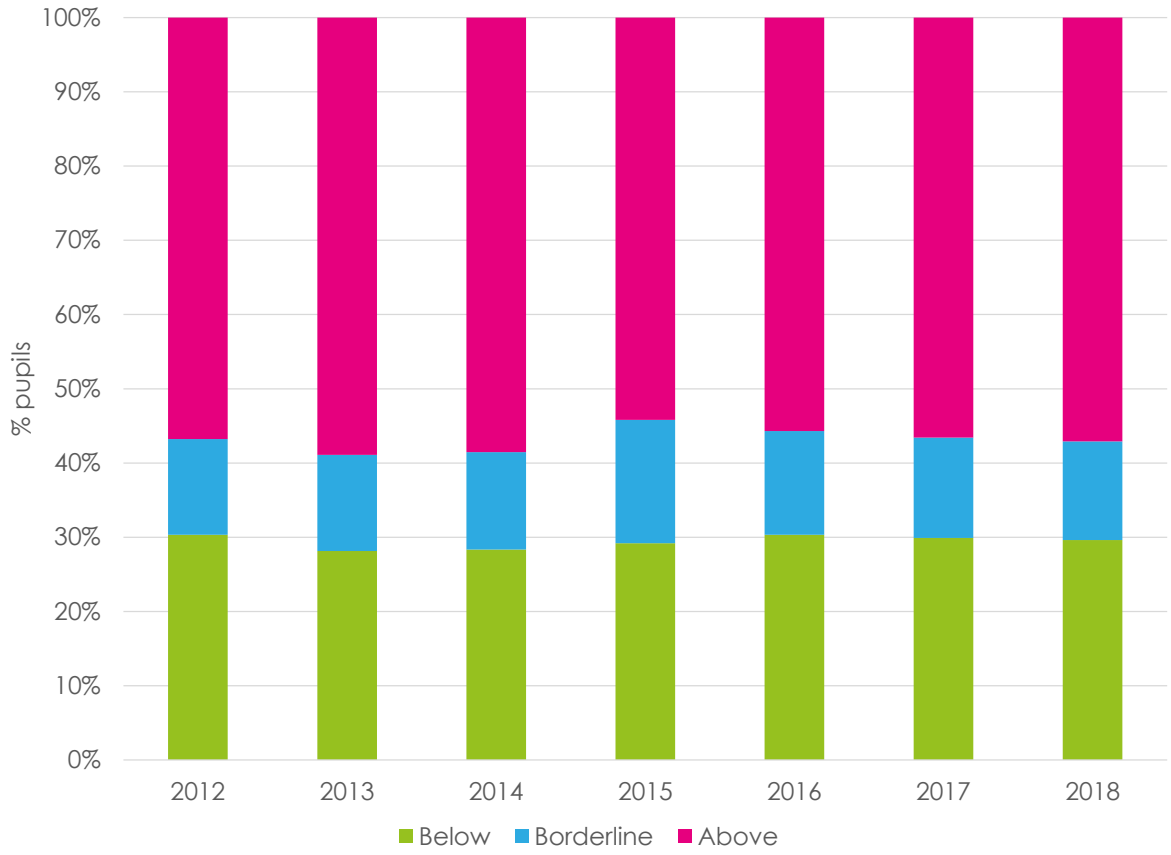
First, we simply directly estimate the effect of the introduction of Progress 8 on three attainment outcomes for disadvantaged pupils, shown in Table 6. These regressions have the same format and control variables as the main regressions in Table 3. We focus on the models that include pre-treatment trends as there are small but statistically significant and negative pre-treatment trends for disadvantaged pupils relative to other pupils in all three outcomes. There were slight increases across all three attainment indicators among disadvantaged pupils following the reform, in particular there was an increase in 0.010 SD in EM points.

Secondly, we provide some insight on the source of that change calculating the change in the impact of disadvantage due to the reform from the policy treatment effects and the differential membership rates in those two groups of disadvantaged pupils. We show this calculation in Table 7. The predicted overall impact is the sum of the two items in row 5, equal to 0.012, nearly the same as in the "reduced form" estimate in Table 6. The interpretation that our model brings is that the improvement for disadvantaged pupils mostly arises because the 'below' group sees the largest improvement in scores and disadvantaged pupils are disproportionately found in this group.

As shown in Table 8, the attainment gap closed by 0.03SD between the pre-reform and post-reform period. The change of 0.012SD from Table 6 represents just under 40% of this change.

Figures and Tables

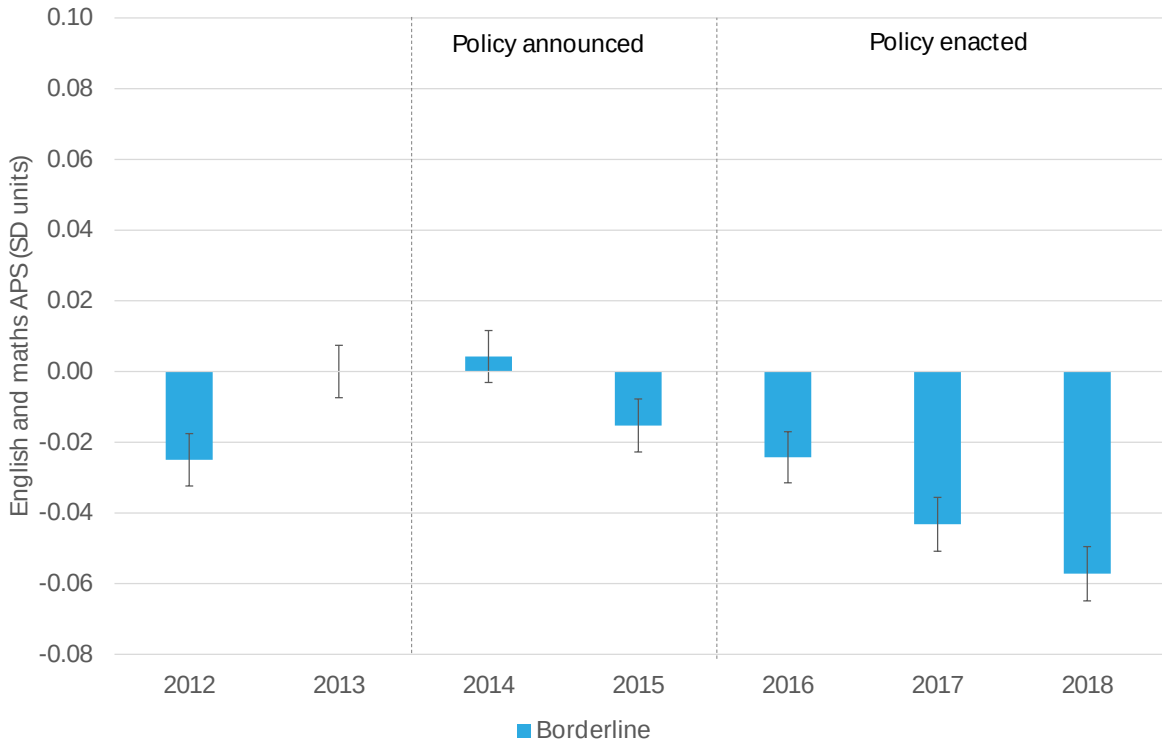
Figure 1: Percentage of pupils in state-funded mainstream schools by group, 2012 to 2018



Notes

This figure shows the proportion of pupils in each cohort who were in the borderline group, i.e. had a 40% to 60% probability of achieving 5 or more GCSEs at grades A*-C including English and maths. See Table 2 for numbers of pupils and schools included

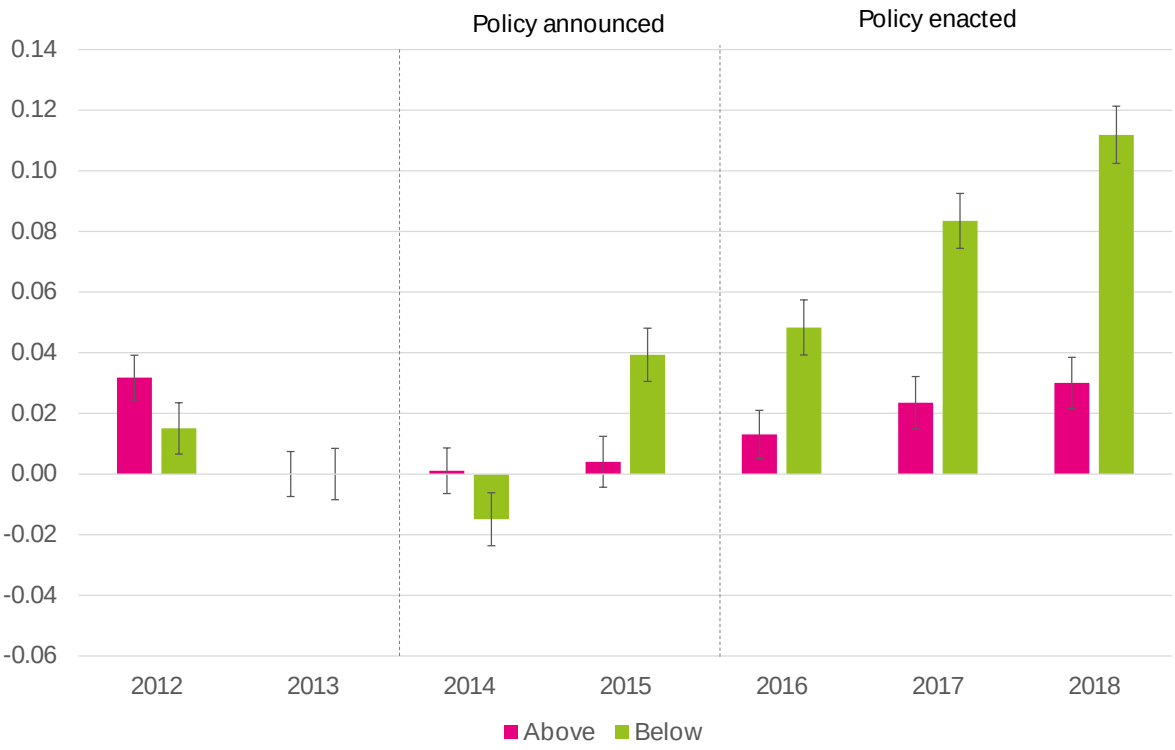
Figure 2: Difference-in-differences estimates for the borderline group by year



Notes

1. The graph shows the coefficients and standard error bands for the δ parameters in equation 3. This is the effect over time on the borderline group of pupils, as defined in the text.
2. The outcome measure is points score in English and maths converted to standard deviation units
3. Standard errors are clustered by school
4. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions of the characteristics with standardized key stage 2 score

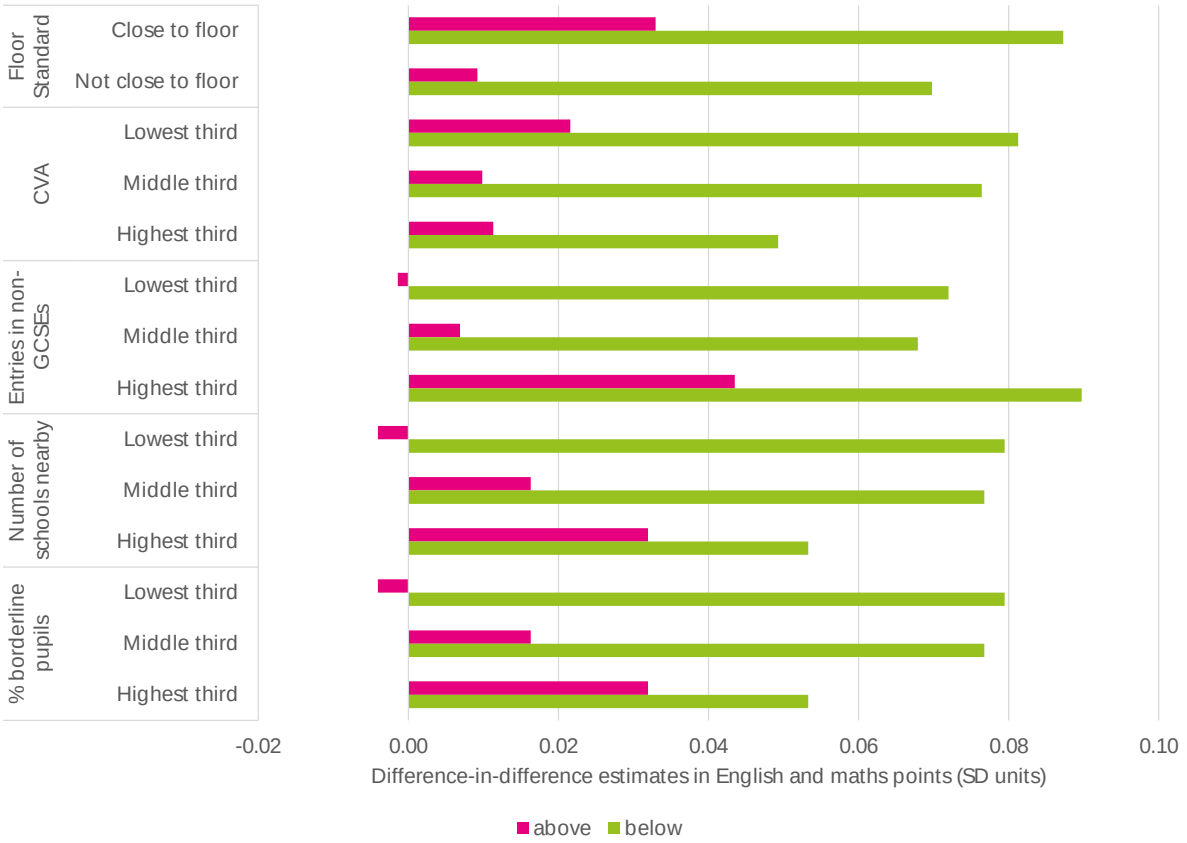
Figure 3: Difference-in-differences estimates for the above- and below-borderline groups of pupils by year



Notes

1. The graph shows the coefficients and standard error bands for the difference-in-differences parameters in equation 2. This is the effect over time on the above- and below-borderline groups of pupils, as defined in the text.
2. The outcome measure is points score in English and maths converted to standard deviation units
3. Standard errors are clustered by school
4. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions of the characteristics with standardized key stage 2 score

Figure 4: Difference in difference estimates interacted with school characteristics, English and maths average point score (SD units)



Notes

1. The bars show difference-in-differences estimates interacted with each of the categories listed in the vertical axis
2. The outcome measure is points score in English and maths converted ito standard deviation units
3. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions of the characteristics with standardized key stage 2 score

Table 1: Means and standard deviations of Key Stage 4 outcomes and pupil characteristics, 2012 to 2018

Variable	Mean	Std. Dev.	Min	Max
English and maths grade	40.14	10.02	0	67.5
5 or more A*-C grades including English and maths	61%	0.49	0	1
Mean GCSE grade	39.00	9.92	0	67.5
Above borderline group	57%	0.50	0	1
Borderline below	14%	0.34	0	1
Below borderline group	29%	0.46	0	1
Standardised KS2	0.04	0.95	-3.5	3.5
Female	50%	0.50	0	1
Free school meals	13%	0.34	0	1
English as an additional language	12%	0.33	0	1
Month of birth	6.52	3.49	1	12

Notes:

N= 3,579 thousand pupils in 3,423 schools

In this table, a unit is a pupil, and the numbers for each school are national averages based on all pupils at the end of Key Stage 4 between 2012 and 2018

Table 2: Selected percentiles of school-level percentages of borderline pupils, 2012-2018

Percentile	2012	2013	2014	2015	2016	2017	2018
10th	8%	8%	8%	10%	9%	8%	8%
25th	11%	11%	11%	13%	12%	11%	11%
50th	13%	13%	13%	16%	14%	14%	14%
75th	16%	16%	16%	20%	17%	17%	16%
90th	18%	18%	18%	25%	19%	19%	19%
Number of schools	3,006	3,008	3,020	3,055	3,087	3,128	3,156
Number of pupils (thousands)	526.8	536.9	523.4	516.9	502.1	488.3	484.3

Notes:

In this table, a unit is a school.

Table 3: Key Parameter Estimates from Headline Models

	1		2		3		4		5	
	b	se(b)	b	se(b)	b	se(b)	b	se(b)	b	se(b)
above	0.765**	0.004	0.686**	0.003	-0.032**	0.002	-0.031**	0.002	-0.031*	0.008
below	-0.766**	0.003	-0.747**	0.003	-0.018**	0.002	-0.022**	0.002	-0.022	0.015
reform	0.031**	0.004			-0.011**	0.003				
Interaction of above and reform	-0.003	0.003	-0.008*	0.003	0.013**	0.003	0.012**	0.003	0.012	0.008
Interaction of below and reform	0.079**	0.003	0.077**	0.003	0.069**	0.003	0.070**	0.003	0.070*	0.020
Number of pupils (thousands)	3579		3579		3579		3579		3579	
Number of schools	3243		3243		3243		3243		3243	
R-squared	0.44		0.50		0.64		0.66		0.66	
pupil covariates	No		No		Yes		Yes		Yes	
school*year fixed effects	No		Yes		No		Yes		Yes	

* significant at the 5% level, ** significant at the 0.1% level

Notes

1. The outcome measure is points score in English and maths converted to standard deviation units
2. Standard errors are clustered by school except column (5) which uses 2-way clustering by school and year
3. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions of the characteristics with standardized key stage 2 score

Table 4: Other outcomes

	1 AC5EM (sd units)		2 Mean GCSE (sd units)	
	b	se	b	se
above	0.224**	0.003	-0.018**	0.002
below	-0.249**	0.003	0.005*	0.002
Interaction of above and reform	-0.010*	0.004	0.005*	0.002
Interaction of below and reform	0.041**	0.004	0.048**	0.003
Number of pupils (thousands)		3579		3579
Number of schools		3243		3243
R-squared		0.47		0.63

* significant at the 5% level, ** significant at the 0.1% level

Notes

1. Outcomes are converted to standardized units – column (1) uses a binary indicator of whether a pupil achieved 5 GCSEs including English and maths and column (2) uses the mean grade in all GCSEs (excluding equivalent qualifications)
2. Standard errors are clustered by school
3. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions between each pupil characteristics and standardized Key Stage 2 score

Table 5: Robustness checks

	1. Shorter Panel		2. Longer panel		3. Excluding early adopters	
	b	se(b)	b	se(b)	b	se(b)
above	-0.031**	0.002	-0.025**	0.002	-0.032**	0.002
below	-0.025**	0.002	-0.026**	0.002	-0.025**	0.002
Interaction of above and reform	0.019**	0.003	0.001	0.002	0.015**	0.003
Interaction of below and reform	0.071**	0.003	0.075**	0.003	0.072**	0.003
Number of pupils (thousands)		3052		4111		3209
Number of schools		3228		3320		2916
R-squared		0.65		0.66		0.65

	4. Wider definition of borderline (30%-70%)		5. Narrower definition of borderline (45-55%)		6. With trend parameters	
	b	se(b)	b	se(b)	b	se(b)
above	-0.012**	0.002	-0.031**	0.002	-0.023**	0.004
below	-0.058**	0.003	-0.016**	0.002	-0.055**	0.004
Interaction of above and reform	0.008**	0.002	0.012**	0.003	0.024**	0.005
Interaction of below and reform	0.081**	0.003	0.062**	0.003	0.024**	0.005
Number of pupils (thousands)		3579		3579		3579
Number of schools		3243		3243		3243
R-squared		0.66		0.66		0.66

* significant at the 5% level, ** significant at the 0.1% level

Notes:

1. This table shows the results of changing the specification in column 4 of Table 3. Column (1) reduces the number of pre-reform years by one, Column (2) adds a further pre-reform year, column (3) fits separate school and year fixed effects, column (4) uses a wider definition to define the borderline group, column (5) uses a narrower definition to define the borderline group, column (6) removes 324 schools which opted into the reform a year early and column (7) adds a linear trend in outcome for both the above and below groups.

2. Outcomes are converted to standardized units – column (1) uses a binary indicator of whether a pupil achieved 5 GCSEs including English and maths and column (2) uses the mean grade in all GCSEs (excluding equivalent qualifications)
3. Standard errors are clustered by school
4. Pupil covariates are: standardized Key Stage 2 score, free school meal eligibility, ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions between each pupil characteristics and standardized Key Stage 2 score

Table 6: Outcomes for disadvantaged pupils

Specification	Parameter	1. EM points		2. 5ACEM		3. Mean GCSE	
		b	se	b	se	b	se
1. Assuming common trends	fsm6	-0.204**	0.001	-0.169**	0.001	-0.238**	0.001
	Interaction of fsm6 and reform	0.015**	0.002	0.004	0.002	0.012**	0.002
	Number of pupils (thousands)		3579		3579		3579
	Number of schools		3243		3243		3243
	R-squared		0.67		0.47		0.64
2. With linear trend parameter	Interaction of fsm6 and year	-0.003**	0.001	-0.013**	0.001	-0.002	0.001
	Interaction of fsm6 and reform	0.011**	0.005	0.015**	0.005	0.025**	0.005
	Number of pupils (thousands)		3579		3579		3579
	Number of schools		3243		3243		3243
	R-squared		0.66		0.45		0.63

** significant at the 1% level

Notes

1. Standard errors are clustered by school*year dummy
2. Outcomes are standardised English and maths scores (EM points), the achievement of 5 or more A*-C grades at GCSE including English and maths (5ACEM) and the mean GCSE grade
3. Pupil covariates are: standardized Key Stage 2 score, disadvantage (FSM6) ethnicity, gender, month of birth, IDACI decile, first language (English/ other) and interactions between each pupil characteristics and standardized Key Stage 2 score

Table 7: Impact estimates for disadvantaged pupils based on main results in Table 3, column 4.

Row	Measure	Above	Below
1	Coefficient	0.012	0.070
2	Percentage of not disadvantaged pupils in respective groups:	63%	24%
3	Percentage of Disadvantaged pupils in respective groups:	40%	45%
4	Difference in percentage (row 2 – row 3)	-0.24	0.21
5	Difference*coefficient (row 4 * row 1)	-0.003	0.015

Table 8: Changes in attainment gap for disadvantaged pupils 2012 to 2018

	Standardised English and maths points		Number of pupils		
Year	Not FSM6	FSM6	Not FSM6	FSM6	Gap
2012	0.13	-0.57	397,930	129,067	-0.70
2013	0.16	-0.51	396,496	140,674	-0.67
2014	0.17	-0.50	386,510	137,076	-0.67
2015	0.17	-0.50	380,304	136,815	-0.67
Pre-reform	0.16	-0.52	1,561,240	543,632	-0.68
2016	0.19	-0.47	367,425	134,965	-0.66
2017	0.19	-0.45	360,450	128,183	-0.64
2018	0.20	-0.44	357,643	126,648	-0.64
Post-reform	0.19	-0.45	1,085,518	389,796	-0.65

Notes

1. FSM6 = pupils eligible for free school meals in the last 6 years. NotFSM6 = all other pupils