Mathematics Mastery evaluation report

FFT Education Datalab: Part of the Education Data Service pilot

September 2019

Contents

1 Executive summary

1.1 Methodology

- This report evaluates the effect of taking part in the Mathematics Mastery project, as measured by the attainment of the expected or higher standards in Key Stage 1 (KS1) maths. Our analysis used pupil-level data from the National Pupil Database (NPD) to compare performance in schools that took part in the project to performance in a group of control schools.
- We looked at the overall effect of taking part in the project, as well as the effect by the length of time a school has been participating in the project.
- Multilevel regression models were fitted to the data, with an indicator to flag whether a school had taken part in the project, or their length of participation. The models were adjusted to take account of pupil-level characteristics.
- Four different models were fitted; one for each outcome (achieving the expected level in KS1 maths and achieving a higher level in KS1 maths) in each outcome year (2017 and 2018).
- We estimated odds ratios for each model, and we also calculated effect sizes and equivalent months of additional progress, according to EEF guidance.
- We also estimated the pooled effects for each outcome, combined across the two outcome years.

1.2 Main findings

- We found positive effects on both outcomes in both outcome years.
- The effect on achieving the expected level of progress, while positive, was not large enough to be equivalent to any additional months of progress in either outcome year. However, at the higher level, the effect was equivalent to two months of additional progress in 2017 and one month in 2018.
- There was no clear trend for effects to increase or decrease the longer a school was part of the project.

1.3 Limitations

- Ideally, from an evaluation perspective, schools would have been randomly assigned to a treated group or a control group. As this was not the case, we constructed a control group of schools similar to the schools that took part in the project, using data from the NPD.
- Creating a control group in this way means that we were unable to control for factors not observed or recorded in our data (such as social class, parental occupation or school funding level).
- Some control schools may have taken part in similar projects. If this improved outcomes in control schools, it may have led to underestimation of effects.
- The effects observed should be considered tentative given these limitations.

2 Introduction

In this report, we evaluated the impact of the Mathematics Mastery project on two outcomes, in 2017 and in 2018. The outcomes we looked at are:

- Expected standard: whether or not a child reaches at least the expected standard in maths by the end of Key Stage 1 (KS1)[1]
- Higher standard: whether or not a child is working beyond the expected level in maths by the end of KS1.

This project was previously evaluated through two randomised control trials, funded by the Education Endowment Foundation, conducted in 2011-13. These trials looked at outcomes for Year 1 and Year 7 students, respectively. They found significant positive effects for both groups; most relevant to this evaluation was the effect on Year 1 students, which was estimated as the equivalent of two months of additional progress.

2.1 Modelling framework

In order to evaluate the impact of the Mathematics Mastery project on the schools that took part, we compared their outcomes to outcomes in a group of control schools. We considered schools that joined the project in one of five years: 2012, 2013, 2014, 2015 or 2016. For schools that joined in 2012-15, outcomes were evaluated in both outcome years. For those that joined in 2016, outcomes were evaluated in 2018 only. There were initially 170 schools for evaluation on 2017 outcomes, and 239 for 2018. However, a number of Mathematics Mastery schools had either recently opened or changed provision shortly before joining the project; as historic data on KS1 attainment was not available for these schools, we were unable to match them to suitable controls and they were not included in the evaluation. Once these schools, and others with substantial amounts of missing data, were excluded, 169 schools remained for evaluation on 2017 outcomes, and 219 on 2018.

We began by constructing a group of control schools that were similar to the treated schools with respect to a set of confounding variables. This is known as a quasi-experimental design; we were aiming to construct a situation that mimics ideal experimental conditions, such as a randomised control trial. We did this using various techniques. This process is discussed in more detail in section two.

The matching used a set of confounding variables that fell into three groups: pupil characteristics, school performance and school characteristics. The variables were:

Pupil characteristics, all related to the year the school joined the project:

- proportion of pupils eligible for free school meals in the last 6 years
- proportion of pupils whose first language was one other than English
- gender ratio (percentage of female students)
- proportion of pupils that lived in one of England's most deprived areas (identified by being amongst the bottom 30% by income deprivation affecting children, or IDACI)
- proportion of pupils with special educational needs

[1]Students who reach the expected standard include: students who meet the expected standard, but did not reach the higher standard AND students who meet both the expected and higher standards.

- proportion of pupils who identified as of white British ethnicity

Previous performance, all related to the three years before the school joined the project:

- proportion of pupils meeting the expected standard in maths during KS1
- average fine grade in KS1 maths [2]

School characteristics, related to the year the school joined the project:

- stage of education: infant, junior, primary or all through

Once a control group was constructed, we fit pupil level models to our data. We used a dummy variable for the length of time that a school had participated in the project and a set of control variables, which allowed us to control for differences between pupils. The variables were the following:

- gender (male / female)
- whether their first language is one other than English
- whether they were eligible for free school meals
- whether they had been eligible for free school meals in the last 6 years
- age, indicated by birth month
- prior attainment during reception

For each model, we obtained an average estimate and confidence intervals by bootstrapping. This involves repeatedly creating a new dataset by taking a random sample from the original, then repeating the analysis using the fresh data. We found bootstrapped estimates for all models using 1,000 iterations.

2.2 Overview of prior performance

Figures 1 and 2 show the distribution of two measures of prior performance: whether students achieved the expected standard in maths during KS1, and the average KS1 maths fine grade. The distributions of both are shown in treated schools and all other mainstream state schools in England, the year before schools joined the project. Schools that took part in the Mathematics Mastery project tended to have had lower maths outcomes than other schools before joining the project. This was the case for both fine grades and the percentage of students achieving the expected level. Schools that joined the project in 2012 and 2013 had particularly low outcomes compared to other schools.

[2] Fine grades were a method of converting national curriculum levels into points to facilitate analysis. National curriculum levels, now obsolete, were used to indicate the level students were expected to reach by a certain point in their education. Levels were numbered 1-8, and within each level there were three sub-levels, A-C, indicating whether students were working at the top, middle or bottom of the level. The levels applied to students in KS1, KS2 and KS3; by the end of KS1, the expected standard was level 2, and students that achieved level 3 were deemed to have reached a higher standard. National curriculum levels were translated into fine grades at Key Stage 1 using the following conversion: 21 points were awarded to level 3, 17 points to level 2A, 15 points to level 2B, 13 points to level 2C, 9 points to level 1 and 3 points to working towards level 1.

Figure 1: Distribution of mean percentage of students achieving the expected level in maths, in treated and all other mainstream state schools in England, by start year

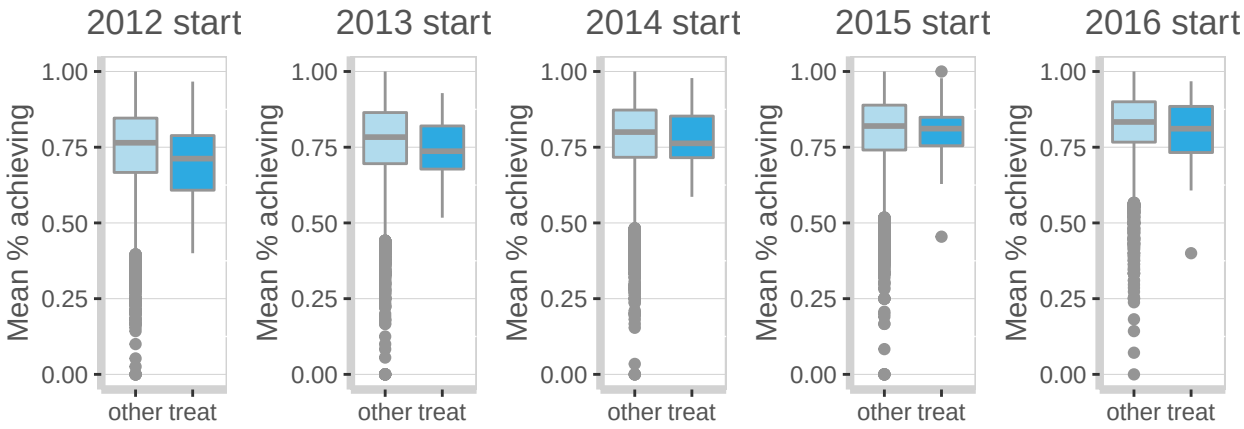


Figure 2: Distribution of average maths fine grade scores, in treated and all other mainstream state schools in England, by start year

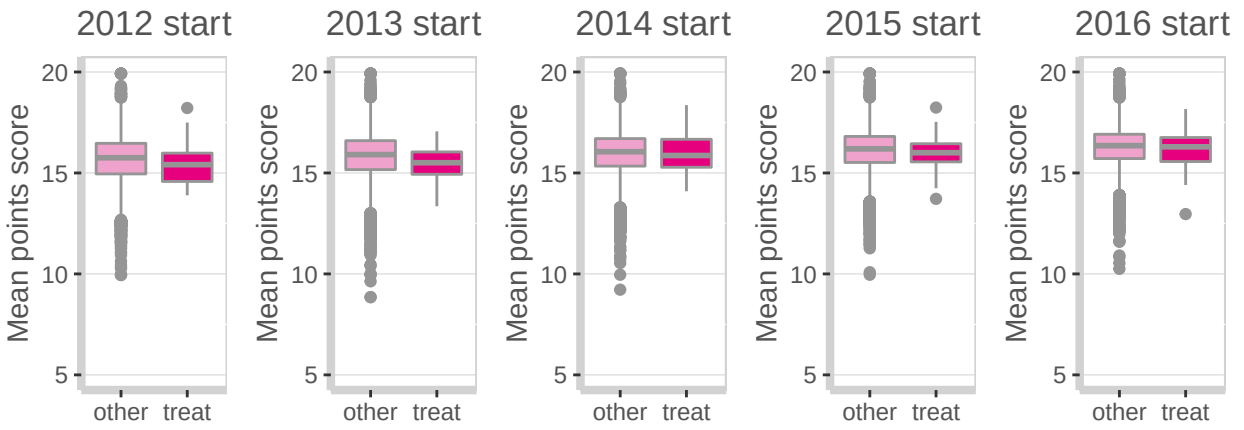
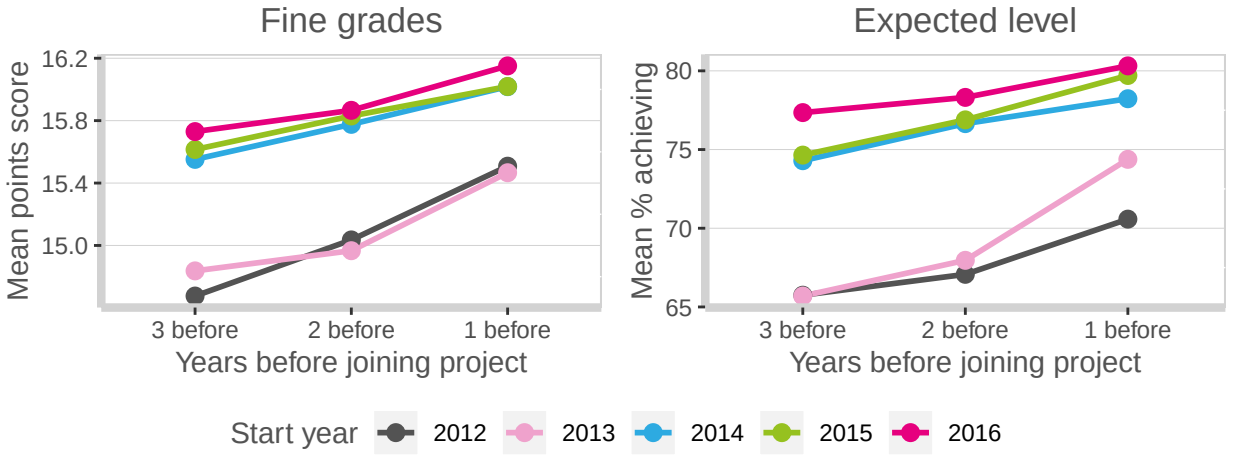


Figure 3: Average performance measures in treated schools during the three years before joining the project



However, although treated schools tended to have relatively low outcomes, they did tend to be improving. We can see from figure 3 that, for all the start years considered, schools' results were improving, on average, for the three years before they joined Mathematics Mastery. It may be the case, then, that any improvements in outcomes after taking part in the project were simply a continuation of that trend, rather than something caused by Mathematics Mastery itself. We can investigate whether this is the case by matching Mathematics Mastery schools to a control group of schools with similar prior performance, and with other similar characteristics. We can then look at whether the Mathematics Mastery schools outperformed these control schools; if they did, that would suggest that the project had a positive effect on the outcomes.

In the next section, we will show how we carried out this matching process.

3 Mitigation of confounding effects

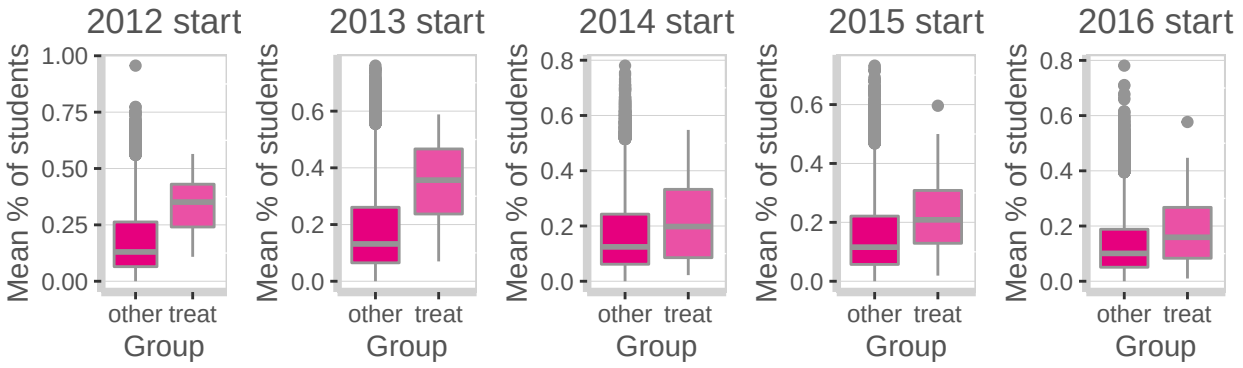
In this section, we will start with an overview of how the Mathematics Mastery schools compared to the potential control schools. We will then discuss how successful our chosen technique was in creating a matched control group.

3.1 Overview

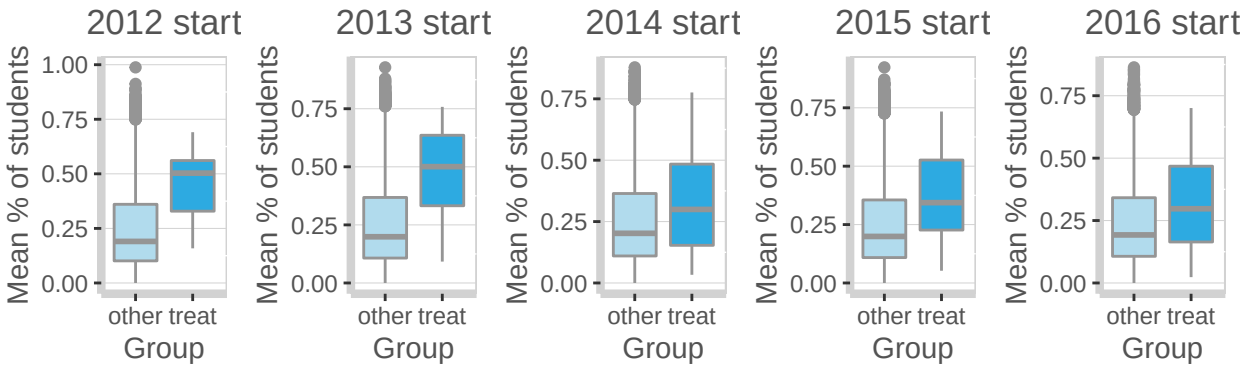
As already discussed (see section 1.2), Mathematics Mastery schools tended to have lower attainment in maths in the few years before they joined the project, but their attainment levels did tend to be improving. They were also unusual in that they had, on average, higher levels of deprivation than other schools. That is, they had higher proportions of FSM students, FSM6 students and students who had lived in one of England's most deprived areas. They also tended to have a higher proportion of EAL students and fewer white British students. This is likely to be because more than half of the treated schools are in the Greater London area; London schools do tend to have this type of profile. From 2014 onwards, schools from outside London began joining the project in greater numbers, and we can see in figure 4 that the pupil characteristics have a wider range after this point.

Figure 4: Distribution of pupil characteristics, in treated and other schools, by start year

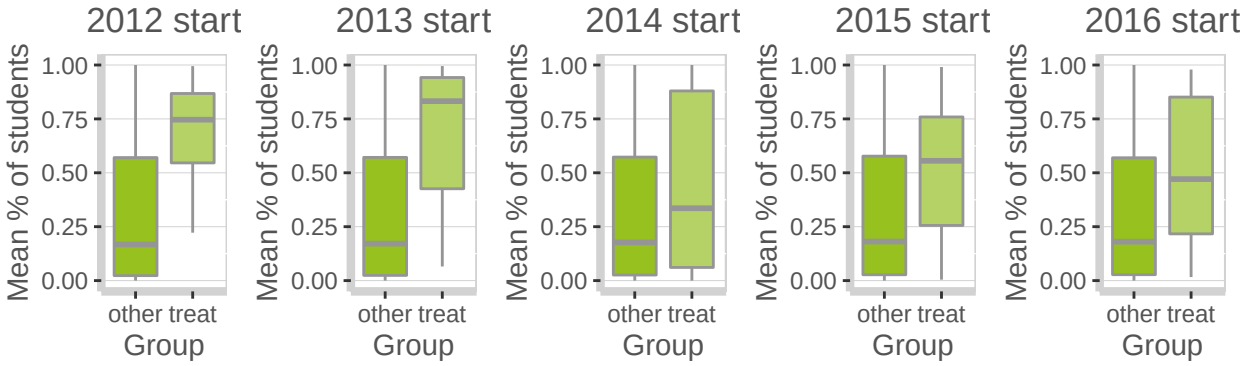
FSM



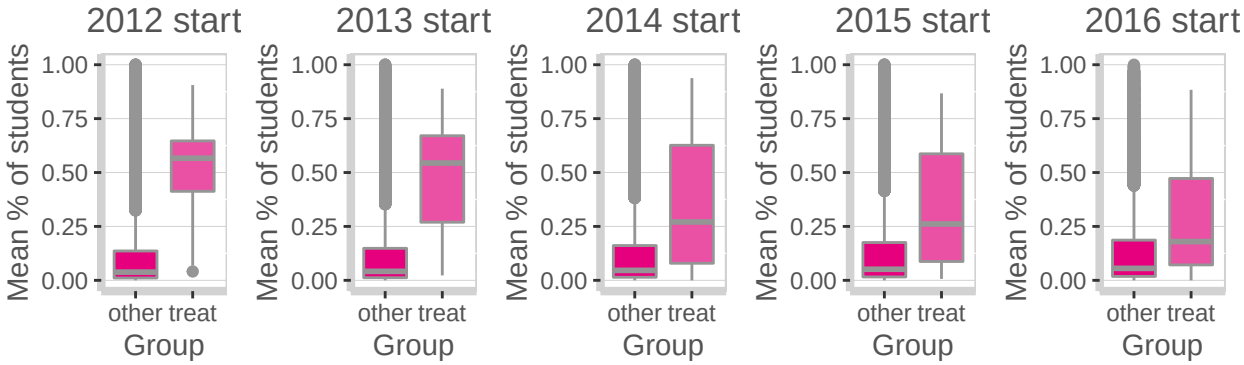
FSM6



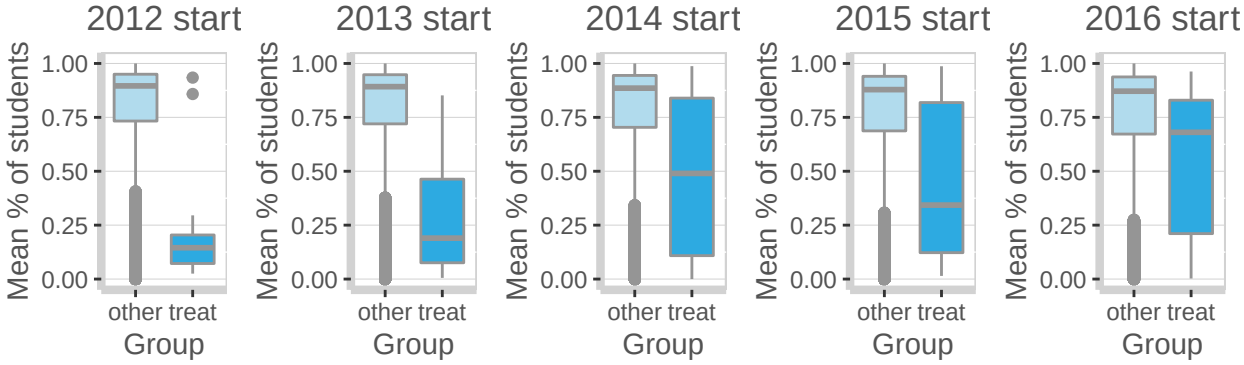
Living in most deprived areas



EAL



White British



NOTE: These charts are boxplots. The line across the middle of the box is the median – the value which exactly half of results fall below, and the other half fall above. The box itself shows the range between the lower quartile – the value below which exactly a quarter of results fall – and the upper quartile – the value above which exactly a quarter of results fall. This definition means that exactly 50% of results are within the range of values covered by the box. The whiskers stretching out from the box reach to the maximum and minimum values.

3.2 Preferred method

Our preferred method for creating matched controls in this case was nearest neighbour matching using propensity scores (NN). This method was chosen because it does not use weighting; adding weighting to a multilevel model with a binomial outcome is problematic. This means that we would have problems using pupil-level data in our model if a method that relies on weighting was used.

We used an additional matching method for schools that joined the project after 2014. As noted in section 3.1 above, from 2014 onwards the characteristics of Mathematics Mastery schools began to change as more schools began to join from around the country; previously they had been concentrated in London. This increased variation in geographical location, and related variation in pupil demographics, made finding matches more difficult. To limit this difficulty, we applied an extra step to the matching process for schools that joined after 2014. This step involved subsetting schools into those in London, and those outside, and only allowing matches with those in the same subset.

Below we show how successful our preferred method was in creating a matched control group.

3.3 Extent of success in creating matched controls

We evaluated how successful our matching methods were by looking at standardised mean differences.

A standardised mean difference is simply a standardised version of the difference between the mean of a variable in the treated group, and the mean in the control group. It is standardised by dividing by the variance of the variable in the treated group. This gives us a rough idea of whether the treated and control groups are similar with respect to the variables considered; if they are similar with respect to all variables, we would conclude that the two groups are well matched. Generally, we consider 'similar' to mean a standardised difference of 0.2 or less.

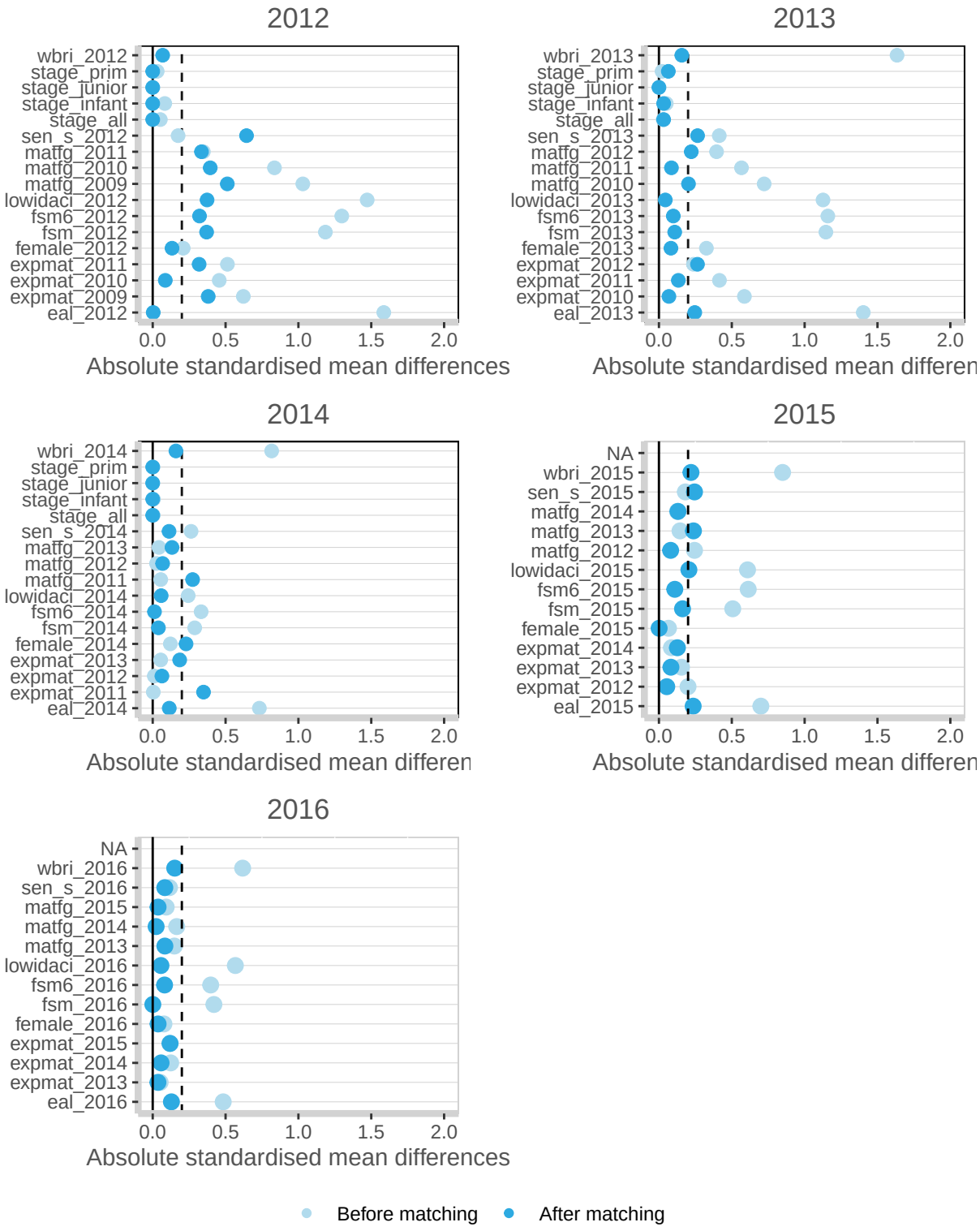
We can visualise how well our control group is matched to our treated group using a type of graph known as a loveplot. In figure 5 (overleaf), we present loveplots for our matched group of schools. We can see that this method reduced the majority of standardised mean differences to within the 0.2 threshold (shown on the graphs as a dotted line). There were some difficulties with finding matches in 2012; this is because only 17 schools were assessed from this start year, once schools with missing data had been excluded.

Key to variables shown on loveplots

- **prop.score**: propensity score (Rubin's B)
- **wbri**: proportion of pupils who identified as of white British ethnicity
- **stage**: stage of education: infant, junior primary or all through
- **sens_s**: proportion of pupils with special educational needs
- **matfg**: average points score in KS1 maths, fine grade
- **lowidaci**: proportion of pupils that lived in one of England's most deprived areas (identified by being amongst the bottom 30% by income deprivation affecting children, or IDACI)
- **fsm6**: proportion of pupils eligible for free school meals in the last 6 years
- **fsm**: proportion of pupils eligible for free school meals
- **female**: gender ratio (percentage of female students)
- **expmat**: proportion of pupils meeting the expected maths standard during KS1

- **eal**: proportion of pupils whose first language was one other than English

Figure 5: Loveplots showing the extent of success in creating a matched control group



4 Results

The results in this section are expressed as odds ratios. These ratios tell us the relative odds of a student attaining either the expected level, or a higher level, in maths at KS1, depending on whether the student attended a school that took part in the Mathematics Mastery project, or a control school. An odds ratio of one would mean that a student from a Mathematics Mastery school has exactly the same odds of attaining the level as a student from a control school. An odds ratio above one means that a student from a Mathematics Mastery school is more likely to attain the level, and an odds ratio of below one means that they are less likely.

We also calculated effect sizes and equivalent months of progress for each outcome. Here, we converted the log odds ratio to the equivalent standardised mean difference using the following formula:

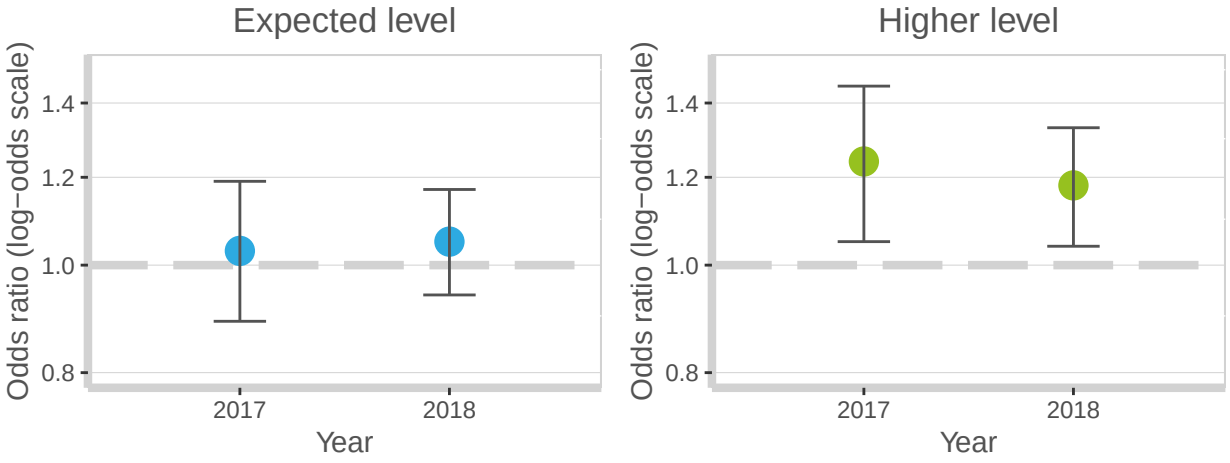
$$d = LogOddsRatio \times \frac{\sqrt{3}}{\pi}$$

Effect sizes were then translated into months of progress using EEF guidelines.

4.1 Overall model

Results from the overall model are summarised in figure 6 below, and given in full, including estimated confidence intervals, in table 1. All of the confidence intervals shown in the results section are at the 95% level, and all results in the tables are given to two decimal places.

Figure 6: Estimated odds ratios for outcome measures, 2017 and 2018



We can see that Mathematics Mastery did appear to have a positive impact on achieving the expected standard, and a higher standard in KS1 maths. The effect on achieving a higher standard was significant; that is, the lower confidence interval was above one. This effect is the equivalent of two months of additional progress in 2017, and one month in 2018.

Table 1: Estimated odds ratios, effect sizes and months of progress for outcome measures in 2017, 2018 and pooled

Type	Year	Outcome	Lower.CI	Estimate	Upper.CI	Months of progress
Odds ratio	2017	Expected level	0.89	1.03	1.19	-
		Higher level	1.05	1.24	1.45	-
	2018	Expected level	0.94	1.05	1.17	-
		Higher level	1.04	1.18	1.33	-
	Pooled	Expected level	0.96	1.04	1.14	-
		Higher level	1.09	1.20	1.32	-
Effect size	2017	Expected level	-0.07	0.02	0.09	0
		Higher level	0.02	0.12	0.20	2
	2018	Expected level	-0.03	0.03	0.09	0
		Higher level	0.02	0.09	0.16	1
	Pooled	Expected level	-0.02	0.02	0.07	0
		Higher level	0.05	0.10	0.15	2

4.2 Models accounting for length of school's participation in the programme

Results are summarised in figures 7 and 8, and full results, including estimated confidence intervals, are given in table 2.

We can see that there was no clear trend for effects to increase or decrease the longer schools are part of the project. Looking at both 2017 and 2018 outcomes, we can see that the most successful group with respect to the second outcome were the schools that joined in 2012 (with participation length of five years in 2017, and six years in 2018).

Figure 7: Estimated odds ratios by outcome measure and length of school participation, 2017

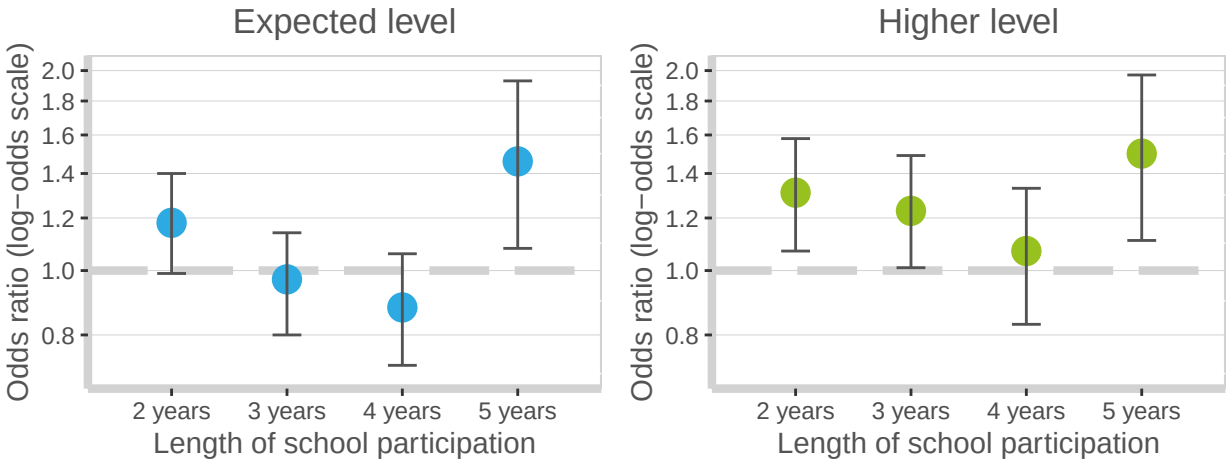


Figure 8: Estimated odds ratios by outcome measure and length of school participation, 2018

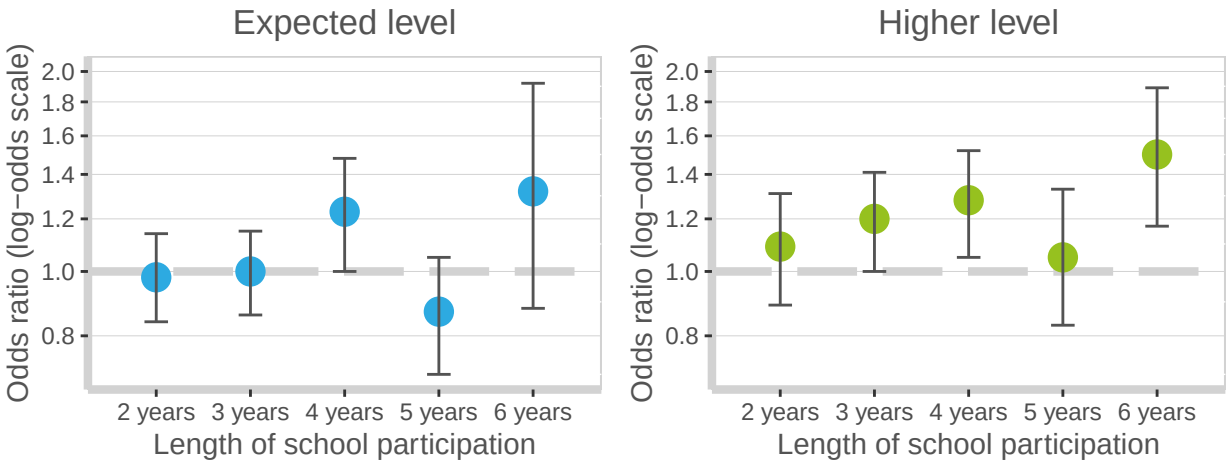


Table 2: Estimated odds ratios, effect sizes and months of progress for outcome measures, 2017 and 2018, by length of school participation

Odds ratios

Length	Year	Outcome	Lower.CI	Estimate
2 years	2017	Expected level	0.99	1.18
		Higher level	1.07	1.31
	2018	Expected level	0.84	0.98
		Higher level	0.89	1.09
3 years	2017	Expected level	0.80	0.97
		Higher level	1.01	1.23
	2018	Expected level	0.86	1.00
		Higher level	1.00	1.20
4 years	2017	Expected level	0.72	0.88
		Higher level	0.83	1.07
	2018	Expected level	1.00	1.23
		Higher level	1.05	1.28
5 years	2017	Expected level	1.08	1.46
		Higher level	1.11	1.50
	2018	Expected level	0.70	0.87
		Higher level	0.83	1.05
6 years	2017	Expected level	0.00	0.00
		Higher level	0.00	0.00
	2018	Expected level	0.88	1.32
		Higher level	1.17	1.50

Effect sizes and months of progress

Length	Year	Outcome	Lower.CI	Estimate	Upper.CI
2 years	2017	Expected level	-0.01	0.09	0.19
		Higher level	0.04	0.15	0.25
	2018	Expected level	-0.10	-0.02	0.07
		Higher level	-0.07	0.04	0.15
3 years	2017	Expected level	-0.12	-0.02	0.07
		Higher level	0.01	0.11	0.22
	2018	Expected level	-0.08	0.00	0.08
		Higher level	0.00	0.10	0.19
4 years	2017	Expected level	-0.18	-0.07	0.03
		Higher level	-0.10	0.03	0.16
	2018	Expected level	0.00	0.11	0.21
		Higher level	0.03	0.13	0.23
5 years	2017	Expected level	0.04	0.20	0.36
		Higher level	0.06	0.21	0.37
	2018	Expected level	-0.20	-0.08	0.03
		Higher level	-0.10	0.02	0.16
6 years	2017	Expected level	0.00	0.00	0.00
		Higher level	0.00	0.00	0.00
	2018	Expected level	-0.07	0.14	0.36
		Higher level	0.09	0.22	0.35

5 Discussion

5.1 Overview

In terms of evaluating the impact of Mathematics Mastery on attainment in KS1 maths, we did find that students that attended a school that took part in the project were slightly more likely to attain both the expected level, and a higher level, in both 2017 and 2018. At the expected level, this effect, while positive, was not significant, and was not large enough to be equivalent to any additional months of progress in either outcome year. However, at the higher level, the effect was equivalent to two months of additional progress in 2017 and one month in 2018. This effect did vary with the length of school participation, although there was not a clear trend.

5.2 Limitations

This evaluation matched treated schools to control schools using observational data from the National Pupil Database (NPD). This type of evaluation is known as a quasi-experimental design. However, ideally, from an evaluation perspective, the project would have been provided to schools as part of a randomised control trial (RCT).

With a quasi-experimental design, there are a number of possible problems. In our analysis, we had to rely on the data in the NPD, but the NPD data is limited. For example, it does not include information about social class, parental occupations or school funding levels. Not accounting for these unobserved variables may introduce bias into our estimates. Using a quasi-experimental design also leaves open the question of how schools were selected to join the project. If there were systemic differences between the Mathematics Mastery schools and control schools - for example, if the project targeted schools in which teachers had low confidence in teaching maths - then these selection effects would pose difficulties to the evaluation.

Some control schools may have taken part in similar projects, or teachers from those schools may have attended training similar to that offered by the Mathematics Mastery project. If this was the case, our analysis would not be an evaluation of the Mathematics Mastery project against no equivalent support, but instead against no support in some cases and other, similar support in the rest. This could lead us to underestimate the effect of the Mathematics Mastery project, assuming that the equivalent support had a positive effect on some control schools' outcomes. We would note, however, that not controlling for this effect may be the relevant analysis as it represents an evaluation of Mathematics Mastery against current conditions, with schools' choices to engage with other projects or training being included in the makeup of controls.

In conclusion, we would be tentative in asserting that the results of this evaluation represent the true size of the Mathematics Mastery project's impact for the reasons outlined above. The ideal evaluation of the project would have come from a fully randomised control trial which would allow for isolation of project participation as a lone variable of interest. As this was not the case, the above results represent the best estimate of the effectiveness of participation in the project that we were able to provide.

6 Appendix: Sensitivity analysis

In this section, we present results obtained from using alternative techniques to mitigate confounding effects. Our preferred method involved using the nearest neighbour (NN) approach to create one-to-one matches, based on propensity scores. We then fit a three-level model to the matched data: students within schools within matched pairs. However, there were some issues with finding good matches for schools that joined the project in 2012.

Here, we used two alternative methods for matching 2012 joiners. We also show results from our preferred model for comparison. Table 3 (overleaf) includes results from the following three models:

- **Model A:** the same model and results as shown in the main results section above, included here for comparison.
- **Model B:** a model that omits 2012 joiners altogether.
- **Model C:** uses an alternative matching technique, coarsened exact matching (CEM) for 2012 joiners.

We can see from table 3 that estimates from the three models were very similar. The estimated months are progress were the same for all three models, with two exceptions: in one case, model C estimated two months of progress while model A estimated three, and vice versa in the other case. However, in every case in which results were obtained, all three models agreed on whether effects are positive or negative and whether they were significantly different from zero.

This sensitivity analysis supports our conclusion that there was no clear trend for effects to increase or decrease the longer a school was part of the project.

Table 3: Estimated effect sizes and months of progress for outcome measures, 2017 and 2018, by length of school participation and model

Length	Year	Outcome	- (A)	Est (A)	+ (A)	MP (A)	- (B)	Est (B)	+ (B)	MP (B)	- (C)	Est (C)	+ (C)	MP (C)
2 yrs	2017	Expected	-0.01	0.09	0.19	1	-0.01	0.09	0.19	1	-0.01	0.09	0.19	1
		Higher	0.04	0.15	0.25	2	0.00	0.15	0.25	2	0.03	0.14	0.25	2
	2018	Expected	-0.10	-0.02	0.07	0	-0.10	-0.01	0.07	0	-0.10	-0.01	0.07	0
		Higher	-0.07	0.04	0.15	0	-0.07	0.04	0.15	0	-0.06	0.04	0.15	0
3 yrs	2017	Expected	-0.12	-0.02	0.07	0	-0.12	-0.02	0.08	0	-0.12	-0.02	0.08	0
		Higher	0.01	0.11	0.22	2	0.00	0.11	0.22	2	0.00	0.11	0.21	2
	2018	Expected	-0.08	0.00	0.08	0	-0.08	0.00	0.09	0	-0.08	0.00	0.08	0
		Higher	0.00	0.10	0.19	2	-0.01	0.10	0.19	2	0.00	0.10	0.20	2
4 yrs	2017	Expected	-0.18	-0.07	0.03	<0	-0.02	-0.07	0.04	<0	-0.18	-0.07	0.04	<0
		Higher	-0.10	0.03	0.16	0	-0.01	0.03	0.16	0	-0.11	0.03	0.15	0
	2018	Expected	0.00	0.11	0.21	2	0.00	0.11	0.22	2	0.00	0.11	0.22	2
		Higher	0.03	0.13	0.23	2	0.02	0.13	0.23	2	0.03	0.13	0.23	2
5 yrs	2017	Expected	0.04	0.20	0.36	3	NA	NA	NA	NA	-0.05	0.22	0.47	3
		Higher	0.06	0.21	0.37	3	NA	NA	NA	NA	-0.12	0.12	0.35	2
	2018	Expected	-0.20	-0.08	0.03	<0	-0.19	-0.08	0.03	<0	-0.19	-0.08	0.03	<0
		Higher	-0.10	0.02	0.16	0	-0.10	0.02	0.15	0	-0.10	0.03	0.16	0
6 yrs	2017	Expected	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		Higher	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	2018	Expected	-0.07	0.14	0.36	2	NA	NA	NA	NA	-0.08	0.24	0.58	3
		Higher	0.09	0.22	0.35	3	NA	NA	NA	NA	0.10	0.26	0.42	3

Key -: Lower confidence limit **Est:** Estimate **+:** Upper confidence limit **MP:** Months of progress

Note: NAs appear where estimates could not be given. For example, there are no estimates for six years' participation for 2017 outcomes, as the first year of the project was 2012; the longest possible length of school participation in 2017 was five years.