Magic Breakfast 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 Magic Breakfast project, as measured by Key Stage 2 (KS2) reading and maths scores. We used pupil-level data from the National Pupil Database (NPD) to compare performance in schools that took part in the 2015 randomised control trial (RCT) of the project to performance in a group of control schools.
- Multilevel regression models were fitted to the data, with an indicator to flag whether a school had taken part in the project. The models were adjusted to take account of pupil-level characteristics.
- Four different models were fitted; one for each outcome (KS2 reading and KS2 maths) in each outcome year (2017 and 2018).
- We estimated treatment effects for each model, and we also calculated effect sizes and equivalent months of additional progress, according to guidance from the Education Endowment Foundation (EEF).

1.2 Main findings

- We did not find any significant effect on KS2 reading or maths scores in 2017. Although we did find positive effects, equivalent to two months' additional progress, for both outcomes, these effects cannot be considered significant; that is, we cannot be confident that these effects were due to the impact of the project.
- However, we did find a significant positive effect on both outcomes in 2018. Based on this finding, we would expect students in a school that took part in the Magic Breakfast RCT to attain scores 1.41 higher in KS2 reading and 1.15 higher in KS2 maths than students in a control school. Alternatively, this could be expressed as an effect size of 0.235 and 0.223 respectively; both effect sizes are the equivalent of three additional months of progress.

1.3 Limitations

- Ideally, from an evaluation perspective, schools would have been randomly assigned to a treated group or a control group. Although this was done for the initial RCT in 2015, we were unable to use the control group from the original trial as controls for this evaluation, as the majority of the control schools went on to take part in the Magic Breakfast project after the trial.
- Therefore, we constructed a control group of schools similar to the schools that took part in the original RCT, 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 evaluate the impact of the Magic Breakfast project. This project works with deprived primary schools, supporting the provision of free breakfast clubs for pupils. Deprived schools are defined as those in which at least 35% of pupils are eligible for free school meals (FSM), and / or at least 50% have been eligible in the last six years (FSM6).

The Magic Breakfast project was evaluated as part of an EEF randomised control trial (RCT) in 2015. For this report, we recreated the RCT using a non-experimental design, focusing on the project's impact on Key Stage 2 (KS2) reading and maths scores. We then considered longer term effects by looking at whether schools that took part in the RCT had improved outcomes in both 2017 and 2018, two and three years after the RCT, respectively.

2.1 Modelling framework

In order to evaluate the impact of the Magic Breakfast project on the schools that were in the treated group during the 2015 RCT, we compared their outcomes to outcomes in a group of control schools. The RCT, of course, included a group of control schools. However, it isn't possible to use these schools as controls in our current evaluation. This is because all of these schools joined the Magic Breakfast project after the RCT was complete. Instead, we need a control group made up of schools that have never taken part in the project.

In the original RCT, there were 53 schools in the treated group, and 53 in the control group. Here, we used data from only 49 of the 53 treated schools. This is because three of the treated schools were infant schools, so do not teach up to KS2. The fourth was a secondary school until 2012, before converting to an all through, and did not have any students in KS2 until 2018. The RCT also looked at outcomes in KS1, which is why these schools were included, but we aren't considering those outcomes here.

We began by creating a pool of potential control schools, consisting of all state-funded mainstream primary schools that would have been eligible to join the project, but did not. From this pool, we constructed a group of control schools that are similar to the treated schools with respect to a set of confounding variables. This is known as a quasi-experimental design; we are aiming to construct a situation that mimics ideal experimental conditions, such as a randomised control trial. We used propensity scores to carry out nearest neighbour matching to create a matched control group.

The confounding variables that we used fell into two groups: pupil characteristics and school performance. The variables were:

Pupil characteristics, all related to the year of the RCT (2015):

- 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
- proportion of pupils who identified as of white British ethnicity

Previous performance, all related to the three years before the RCT:

- average points score in KS1
- proportion of pupils meeting the expected standard during KS2
- average points score in KS2 English
- average points score in KS2 maths

Once a control group was constructed, we fitted pupil level models to our data. These multilevel regression models accounted for the structure of our data: pupils within schools. We included a set of control variables in each model; these allow us to control for differences between pupils. The control variables were the following:

- prior attainment at KS1
- 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

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.

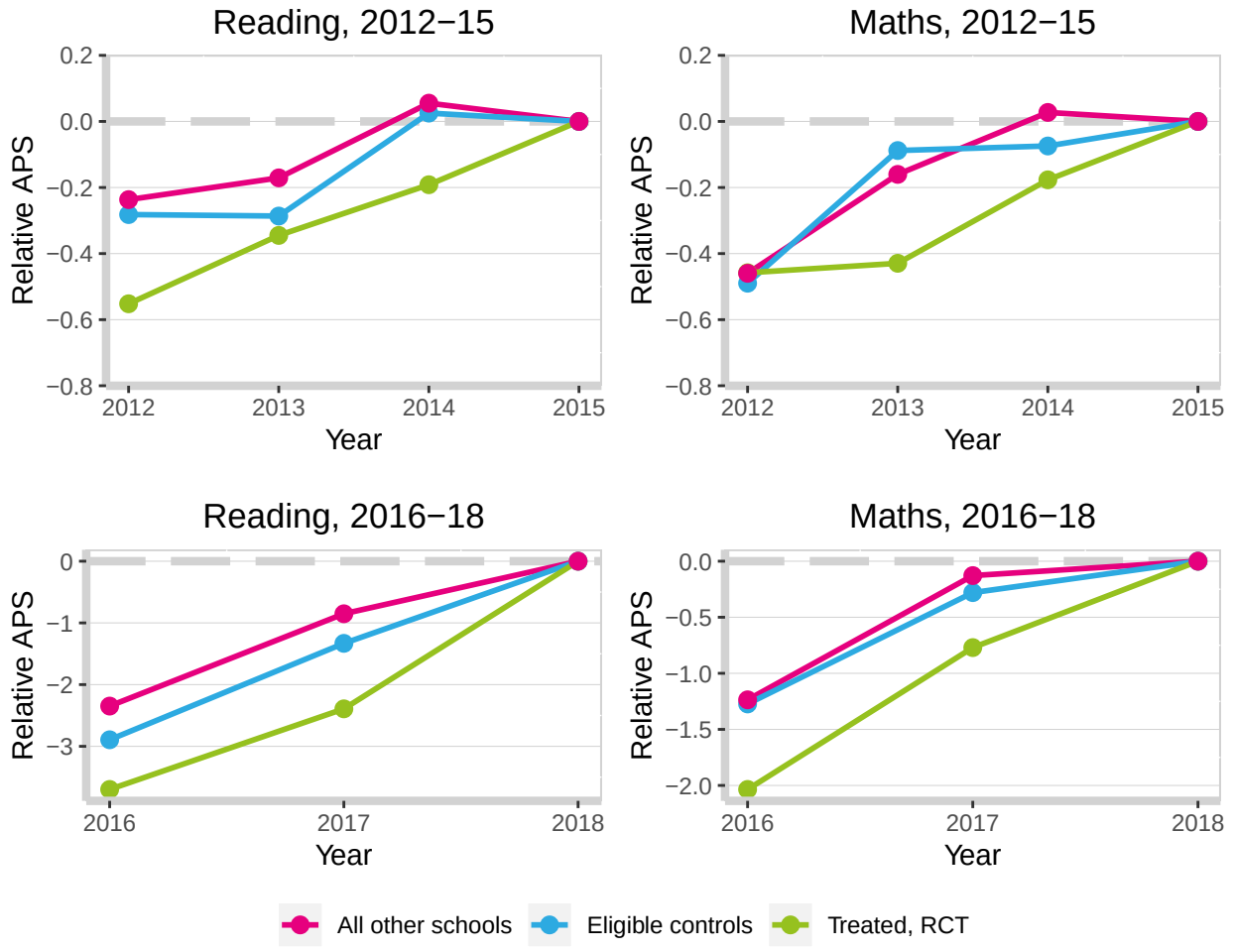
2.2 Overview of outcome measure

We looked at two outcome measures: KS2 attainment in maths, and KS2 attainment in reading. Figure 1 gives an overview of how the mean outcomes in treated schools compared to other schools in the year of the RCT, and the three years before. It is not surprising, considering that only deprived schools were eligible for the project, to see that outcomes in both treated schools and eligible controls were below those in other schools.

Comparing the outcome measures prior to the RCT with those from 2016 onwards is slightly more complicated, as the way KS2 attainment was reported changed in that year, moving from a raw score, with typical values of roughly 25-30, to a scaled score. With the scaled scores, 100 represents the expected standard, and 110 a high standard. Figure 1 includes separate graphs for 2016-18.

We can see from figure 1 that outcomes in treated schools for both reading and maths did improve from 2014 to 2015, and continued to improve from 2016-18. In fact, between 2016 and 2018 treated schools had slightly better outcomes than eligible control schools. However, it does also appear that both outcomes were improving in treated schools before the RCT began. It may be the case, then, that any improvements in outcomes after taking part in the project are simply a continuation of that trend, rather than something caused by the Magic Breakfast project itself. Using matching and weighting techniques, as well as including control variables in our models, will allow us to control for these trends and help us to isolate the effect of the project.

Figure 1: School level outcome measures, 2012-18. The 2012-2015 graphs show the difference between results in 2015 and those in previous years. The 2016-18 graphs show the difference between results in 2018 and those in 2016-17



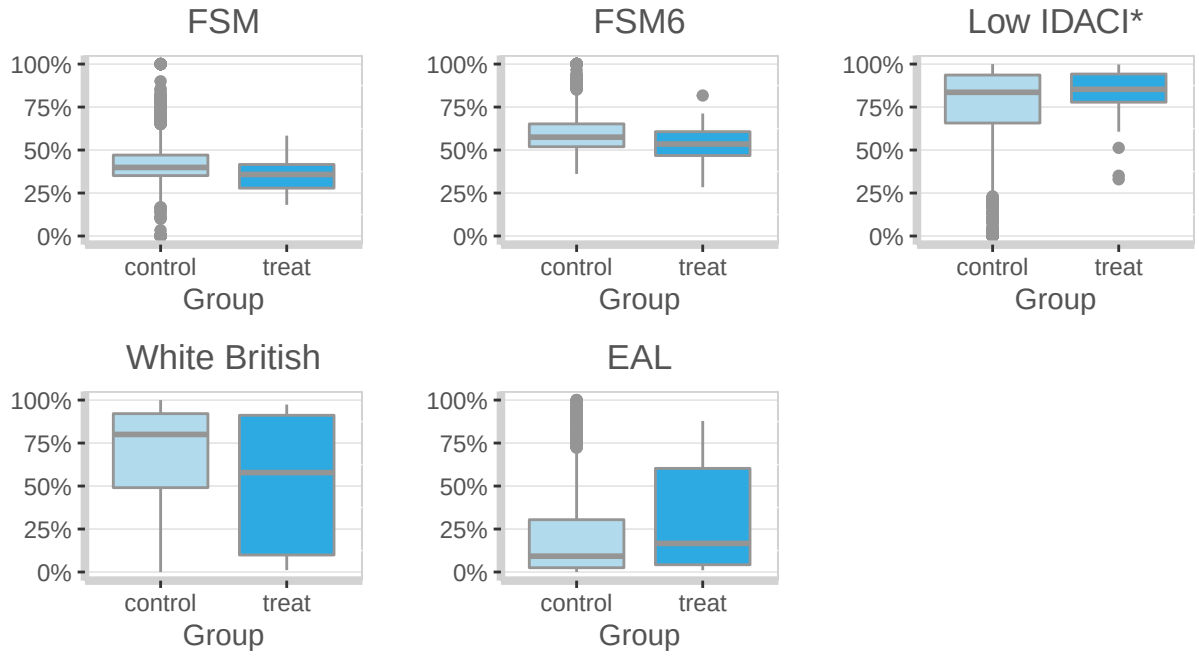
3 Mitigation of confounding effects

3.1 Overview

The group of potential control schools included all mainstream state schools in England that met the criteria for taking part in the Magic Breakfast project, but had not taken part in the project at any time. The criteria for taking part in the project is that at least 35% of pupils are eligible for free school meals (FSM), and / or at least 50% have been eligible in the last six years (FSM6).

Even with only eligible schools included in the potential control group, the treated schools were still relatively deprived, as measured by FSM, FSM6 and the proportion of 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 probably because they were concentrated in urban areas, particularly Greater London. As discussed above (see section 1.2), the treated schools did tend to have slightly lower outcomes than the potential control schools. There were no large differences in the distribution of the proportion of female or SEN students between the treated schools and potential controls.

Figure 2: Distribution of pupil characteristics, treated and potential control schools



*Pupils who have lived in one of the most deprived areas, identified by being amongst the bottom 30% by income deprivation affecting children, or IDACI

3.2 Creating a matched control group

We matched treated schools to control schools using propensity scores. A school's propensity score is its probability of being in the treated group given its values for the confounding variables considered. This is usually estimated by fitting a logistic regression model to a dataset which includes all treated schools and a group of potential control schools. Each treated school is then paired with a control school with the closest possible propensity score.

We can assess how well-matched a control group is in a number of ways. Here, we will look at standardised mean differences and variance

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 calculated using this formula:

$$MD_S = \frac{\mu_{tr} - \mu_{con}}{\sigma_{tr}}$$

... where $\mathrm{MD_S}$ is the standardised difference for a variable, μ_{tr} is the mean of that variable in the treated group, μ_{con} is the mean of that variable in the control group, and σ_{tr} is the variance of that 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 have considered 'similar' to mean a standardised difference of 0.2 or less. This is a common approach.

We can visualise covariate balance by using loveplots, which show the standardised mean differences for each variable, both before and after matching, and the 0.2 threshold as a dotted line.

Variance ratios can be used as another way of assessing balance, in addition to standardised mean differences. They compare the variance of the treated and control group; a variance ratio of one would indicate identical variance. If a covariate has a low standardised mean difference, and a variance ratio close to one, that suggests its distribution is very similar in the treated group and in the control group. We would consider a variance ratio below 2 to be reasonable.

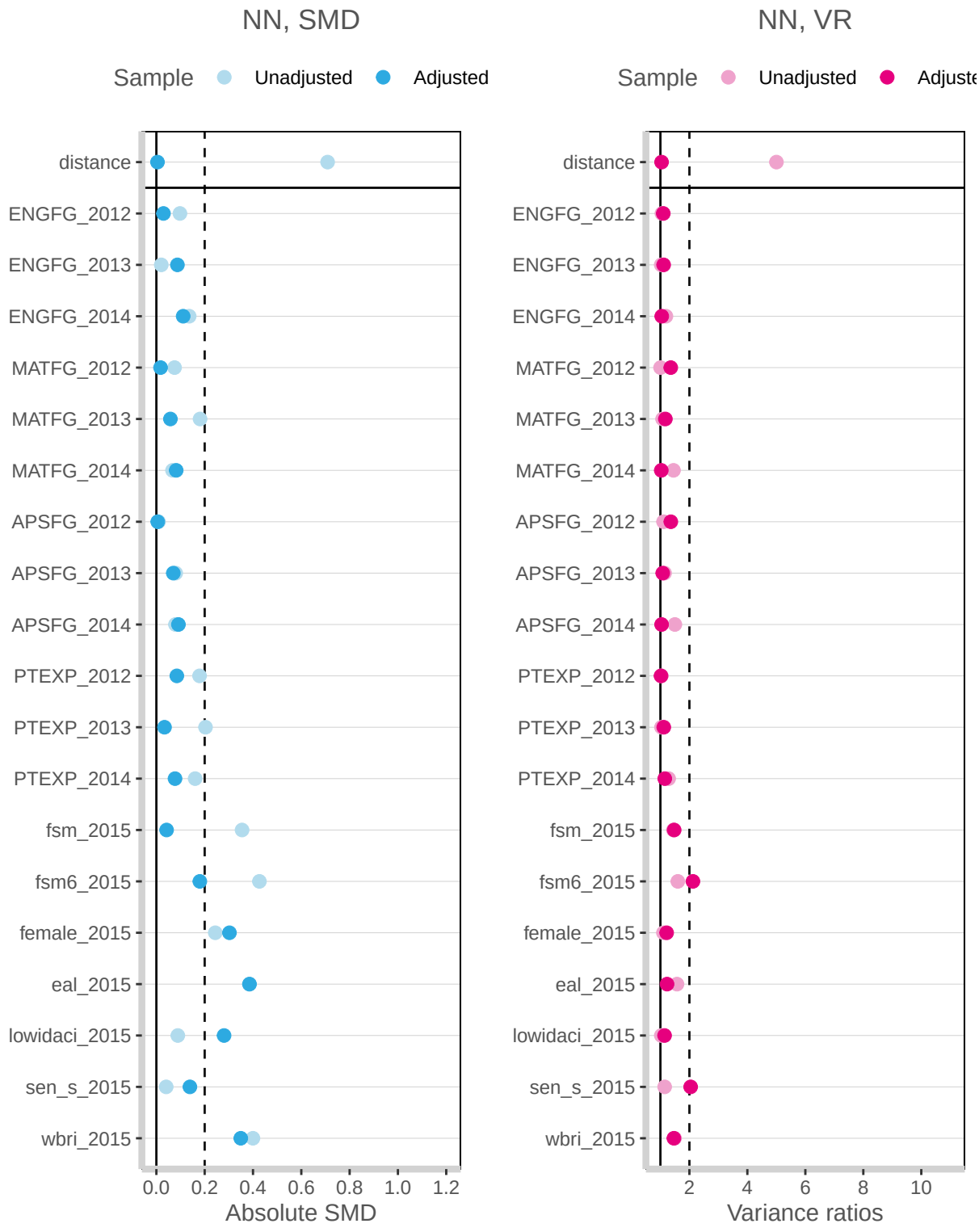
Like standardised mean differences, variance ratios can be effectively visualised using loveplots. In figure 3, we present loveplots of both balance measures. These plots also include information on Rubin's B; that is, the absolute standardised mean difference in propensity score.

Another way of evaluating balance between treated and control groups is to consider the extent of common support. Common support is generally used to refer to the overlap in propensity scores between the two groups. Control schools with propensity scores outside the range of propensity scores among treated schools, or vice versa, would be said not to have common support.

Figure 4 shows the region of common support for the treated and untreated schools before and after matching.

From figures 3 and 4, we can see that the control group was generally well-matched to the treated group, although there was one variable which is over our threshold of 0.2 for standardised mean differences, and three over the threshold for variances ratios. However, this is not too much of a cause for concern given the good matching elsewhere.

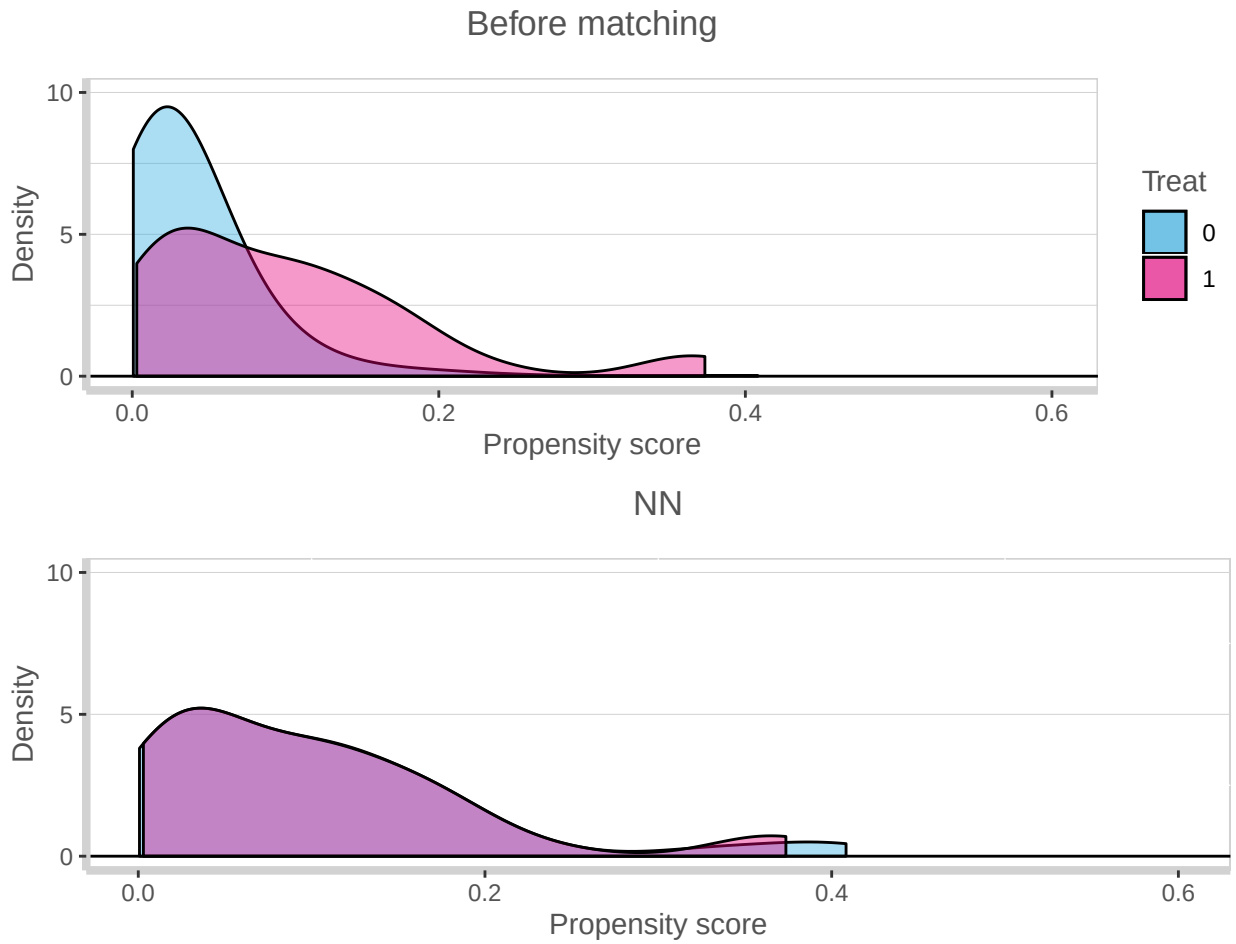
Figure 3: Loveplots showing the standardised mean differences (SMD) and variance ratios (VR) before and after matching



Key to variables shown on figure 3

- **distance**: Rubin's B: propensity score
- **wbri**: proportion of pupils who identified as of white British ethnicity
- **sens_s**: proportion of pupils with special educational needs
- **PTEXP**: proportion of pupils meeting the expected standard during KS2
- **MATFG**: average points score in KS2 maths
- **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)
- **ENGFG**: average points score in KS2 English
- **eal**: proportion of pupils whose first language was one other than English
- **APSFG**: average points score in KS1

Figure 4: Common support, before and after matching / weighting, by matching method



3.3 Success in recreating RCT results

We fit models on 2015 outcomes, using the control groups created by each method. Figure 5 and tables 1 and 2 below show how the estimated effect sizes from these models compare with the estimates from the original RCT. The graphs include a point estimate for each effect size, with the bars representing the upper and lower 95% confidence levels. We can see that the point estimates and confidence intervals are similar, which suggests that we have been successful in constructing an appropriate control group, comparable to that used in the RCT.

Figure 5: Estimated effect sizes on KS2 reading and maths in 2015, by method

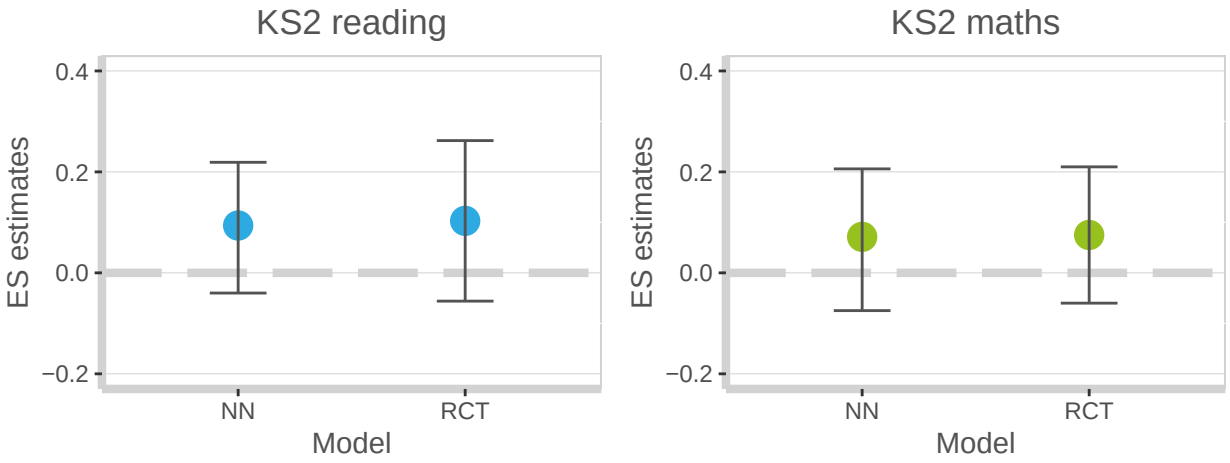


Table 1: Reading effect size estimates by method, 2015

Model	Lower CI	Estimate	Upper CI	Months of progress
NN	-0.04	0.09	0.22	1
RCT	-0.06	0.10	0.26	2

Table 2: Maths effect size estimates by method, 2015

Model	Lower CI	Estimate	Upper CI	Months of progress
NN	-0.07	0.07	0.21	1
RCT	-0.06	0.08	0.21	1

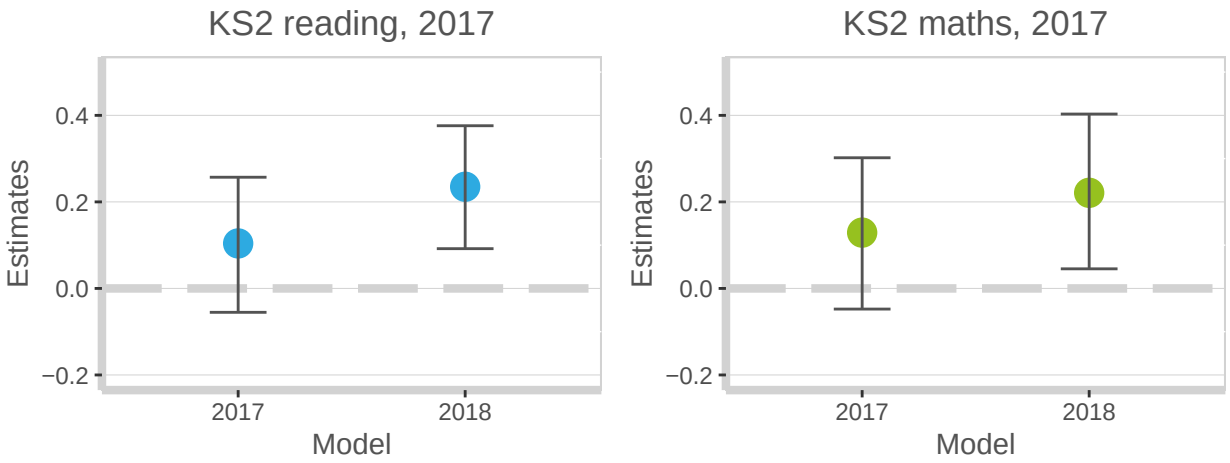
4 Results

4.1 Overall results

We found no significant effects on KS2 reading or maths in 2017, but we did find significant positive effects on both outcomes in 2018. Although we found positive effects on both reading and maths in both 2017 and 2018, the confidence intervals for all estimates in 2017 include zero, these effects cannot be considered significant.

Results are summarised in figure 5, and full results, including estimated treatment effects and estimated effect sizes, are given in tables 1 and 2. All of the confidence intervals shown in the results section are at the 95% level, and all results in the tables are given to three significant figures.

Figure 6: Effect size estimates, reading and maths, by method



Effect sizes were calculated by dividing the estimated treatment effect by the residual standard deviation, using the following equation:

$$d = \frac{t}{\sqrt{v_{res}}}$$

. . . where d is the estimated effect size, t is the estimated treatment effect and v_{res} is the residual variance in the outcome measure. Residual variance is the variance remaining after controlling for the fixed effects included in our models.

Table 3: Reading treatment effect and effect size estimates by year

Year	Method	Lower CI	Estimate	Upper CI	Months of progress
2017	Treatment effect	-0.3280	0.621	1.530	-
	Effect size	-0.0552	0.104	0.257	2
2018	Treatment effect	0.5540	1.400	2.230	-
	Effect size	0.0919	0.235	0.376	3

Table 4: Maths treatment effect and effect size estimates by year

Year	Method	Lower CI	Estimate	Upper CI	Months of progress
2017	Treatment effect	-0.2520	0.670	1.580	-
	Effect size	-0.0477	0.129	0.302	2
2018	Treatment effect	0.2350	1.140	2.060	-
	Effect size	0.0453	0.221	0.403	3

5 Discussion

5.1 Overview

In terms of evaluating the impact of the Magic Breakfast project on KS2 reading and maths in 2017 and 2018, we found positive effects. In 2017, these effects were equivalent to two months' additional progress. However, they were not significant; we did not find any significant effects on the outcomes considered in 2017. In 2018, we found effects that were both positive and significant, equivalent to three months of additional progress.

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). Although this was done for the initial RCT in 2015, we were unable to use the control group from the original trial as controls for this evaluation, as the majority of the control schools went on to take part in the Magic Breakfast project after the trial.

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. Although our control group only included schools that would have been eligible to join the project, there may have been other criteria that made a difference; for example, the practicalities of running the project may have been that schools which are geographically remote were less likely to take part. If there were systemic differences between the treated and control schools then these selection effects would pose difficulties to the evaluation.

Some control schools may have taken part in similar projects. In fact, it was noted in the evaluation of the original RCT that some of the original control group had run similar breakfast clubs even during the trial, so it does seem very likely that at least some of the schools in our constructed control group would also have done so. If this was the case, our analysis would not be an evaluation of the Magic Breakfast 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 Magic Breakfast 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 Magic Breakfast against current conditions, with schools' choices to engage with other projects or run similar breakfast clubs 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 Magic Breakfast 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 an alternative technique to mitigate confounding effects. Our preferred method involves 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. Here, we use covariate balancing propensity scores (CBPS) instead. Rather than creating matched pairs, this method weights the control group so that it is similar to the treated group. We then a two-level weighted regression model: students within schools. As with the NN technique, we use bootstrapping to obtain confidence intervals for our results.

This method was slightly less successful than the NN method in recreating the results of the 2015 RCT, tending to slightly underestimate the results, as shown in tables 5 and 6 below.

Table 5: Reading effect size estimates by method, 2015, with CBPS

Model	Lower CI	Estimate	Upper CI	Months of progress
NN	-0.04	0.09	0.22	1
RCT	-0.06	0.1	0.26	2
CBPS	-0.02	0.07	0.16	1

Table 6: Maths effect size estimates by method, 2015, with CBPS

Model	Lower CI	Estimate	Upper CI	Months of progress
NN	-0.07	0.07	0.21	1
RCT	-0.06	0.08	0.21	1
CBPS	-0.08	0.02	0.12	0

In table 7 and 8 below, we show results for 2017 and 2018 from the CBPS technique, in the same format used in tables 3 and 4 in the results section.

Table 7: Reading treatment effect and effect size estimates by year, CBPS technique

Year	Method	Lower.CI	Estimate	Upper.CI	Months.of.progress
2017	Treatment effect	-0.0760	0.452	0.974	-
	Effect size	-0.0126	0.075	0.162	1
2018	Treatment effect	0.8160	1.270	1.720	-
	Effect size	0.1300	0.210	0.286	3

Table 8: Reading treatment effect and effect size estimates by year, CBPS technique

Year	Method	Lower.CI	Estimate	Upper.CI	Months.of.progress
2017	Treatment effect	-0.02370	0.4720	0.977	-
	Effect size	-0.00452	0.0901	0.189	1
2018	Treatment effect	0.26800	0.8440	1.380	-

Year	Method	Lower.CI	Estimate	Upper.CI	Months.of.progress
	Effect size	0.05060	0.1610	0.266	2

As we might expect from the 2015 results, the CBPS technique gives slightly lower estimates than those obtained from the NN technique. However, both techniques suggest positive effects for both outcomes in both outcome years, with significant positive effects in 2018.

This sensitivity analysis supports our conclusion that the Magic Breakfast project is associated with positive effects on both reading and maths, and that these effects are significant in 2018, but not in 2017.