Evaluation of the impact of the SAM Learning e-learning platform on attainment at GCSE

2011-18

Natasha Plaister

Dave Thomson

education datalab

June 2020

Contents

1. Executive summary

1.1 Methodology

- This report evaluates the effect of SAM Learning's eLearning platform on pupils' attainment at GCSE between 2011 and 2018. Our analysis used pupil-level data from the National Pupil Database (NPD) to compare the performance of pupils who had a SAM Learning account to the performance of a statistically matched comparison group.
- The comparison group was created using inverse probability weighting (IPW) based on a set of confounding variables.
- We looked at the overall effect of taking part in the project, as well as the effect for pupils with low, medium and high levels of engagement with SAM Learning, and for disadvantaged pupils.
- Multilevel regression models were fitted to the data, with an indicator to flag whether a pupil had taken part in the project. Controls were used to account for any differences between the treated and comparison group remaining after the matching process.

1.2 Main findings

- This report was prepared as part of the Education Data Service (EDS) pilot. The dataset analysed consisted of 297,973 pupils from 252 schools who used SAM Learning from 2011-18. The EDS pilot was funded by the Education Endowment Foundation.
- We found significant positive effects for high dosage pupils; that is, for pupils who had spent ten hours or more using their SAM Learning account during Year 11. This effect was the equivalent of between a ninth and a third of a grade per subject.
- We also found evidence of a stronger significant positive effect for disadvantaged pupils, defined as those pupils who were eligible for free school meals. For high dosage disadvantaged pupils, this was the equivalent of between a fifth and more than half a grade per subject.
- We did not find any consistent evidence for an overall significant effect on GCSE point scores; that is, we found no consistent significant effect from having a SAM Learning account, including pupils at all dosage levels and those with zero use.

1.3 Limitations

- Ideally, from an evaluation perspective, pupils would have been randomly assigned to a treated group or a control group. As this was not the case, a comparison group was constructed using data from the NPD.
- Creating a comparison 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 motivation level). For example, pupils in the low dosage group may be those with low motivation, resulting in underestimates of SAM Learning's impact, or pupils in the high dosage group may be those with high motivation, resulting in overestimates. We were also unable to match on historical school performance.
- Some comparison pupils may have taken part in similar projects. If this improved outcomes for comparison pupils, it may have led to underestimation of effects.
- The effects observed should be considered tentative given these limitations.

2. Introduction

SAM Learning is a subscription service that provides schools with the tools to carry out a three wave intervention. This includes giving pupils access to an eLearning platform; each pupil in a subscribing school is provided with an account. In this evaluation, we focus on estimating the impact of SAM Learning on average GCSE point scores for pupils with a SAM Learning account during Year 11, from cohorts of pupils from 2011-18.[1] As well as looking at the overall impact for all pupils with an account, we also evaluated the impact by dosage; that is, by how much time pupils spent using their account. Pupils who used their account for 0-2 hours were classed as low dosage, 2-10 hours as medium dosage, and 10 or more hours as high dosage. [2] Finally, we looked at the impact by dosage on pupils who were eligible for free school meals.

The initial dataset provided by SAM Learning consisted of 344,283 pupils from 272 schools. Of these, 21,164 pupils could not be matched to records in the National Pupil Database; these were pupils for whom no valid unique pupil number (UPN), name or date of birth was given in the dataset. A further 19,147 were found not to have been in Year 11 in the year for which data was supplied. Finally, we excluded any pupils for whom data on prior attainment at Key Stage 2, or GCSE grades, was unavailable, leaving a final dataset for analysis of 297,973 pupils from 252 schools. Table 1 below shows the number of pupils in each cohort from 2011-18.

Table 1: SAM Learning pupils by year, 2011-18

Year:	2011	2012	2013	2014	2015	2016	2017	2018
Pupils:	26,054	27,280	29,548	32,350	33,238	35,295	37,856	41,191

This report was prepared as part of the Education Data Service (EDS) pilot. The EDS pilot was funded by the Education Endowment Foundation to explore the feasibility of using quasi-experimental designs based on observational data to carry out evaluations of education projects.

[1] Throughout this report, years refer to the year in which the academic year finished - that is, 2017 refers to 2016/17, for example.

[2] These levels were chosen as they were used in a previous evaluation by the Fischer Family Trust, carried out in 2012: https://www.samlearning.com/proven-impact/

3. Modelling framework

For this evaluation, we used what is known as a *quasi-experimental design*. This involves comparing the outcomes of pupils who had SAM Learning accounts to a matched comparison group of pupils. This approach mimics what would be done in a formal experiment such as a randomised control trial.

Pupils in the matched comparison group were weighted so that they were similar to those who had SAM Learning accounts with respect to:

Pupil characteristics, all related to the outcome year:

- prior attainment at Key Stage 2
- gender (male / female)
- ethnic group
- whether they had special educational needs
- whether they spoke English as an additional language
- whether they were eligible for free school meals
- the proportion of their school career for which they had been eligible for free school meals

School characteristics

- proportion of pupils eligible for the Pupil Premium
- proportion of pupils with English as an additional language
- average prior attainment at Key Stage 2
- region

SAM Learning offers annual subscriptions and schools are not obliged to take part in the intervention for any fixed period of time. The majority of schools are long-term subscribers, many of which signed up ten or more years ago. There are also some questions over the reliability of the join dates recorded in the dataset provided; in some cases, the join date given is later than the date that the data relates to. For these reasons, we were unable to use historical attainment in the outcome measure as a matching variable, as we usually would.

We used regression models to compare outcomes for SAM Learning pupils to pupils in the matched comparison group. In each case, we used a dummy variable to indicate whether a pupil had taken part in SAM Learning, and we used the characteristics listed above as control variables. This *doubly robust* approach means that our results will remain unbiased if either the model used for matching or the regression model is misspecified.

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 using 1,000 iterations.

4. Mitigation of confounding effects

This section begins with an overview of how the pupils who were signed up to SAM Learning compared to other pupils. We then go on to discuss the technique used to create the comparison group and how successful it was.

From this point onwards, we will refer to pupils who were signed up to SAM Learning as *treated pupils* and all other pupils as *potential comparison pupils*.

4.1 Overview

SAM Learning does not target its support to specific groups of pupils or schools; it is open to any school that chooses to subscribe. Reflecting this policy, the profile of treated pupils is generally similar to that of potential comparison pupils. There were some small but consistent differences in terms of the level of disadvantage; for every cohort since 2013, SAM Learning pupils have been slightly less likely than potential comparison pupils to be eligible for free school meals. For the 2018 cohort, for example, the proportion of treated pupils who were eligible for free school meals was 11.5%, compared to 13.5% of potential comparisons. There were also some regional differences; SAM Learning tended to have fewer subscribing schools based in the North East and Yorkshire and the Humber than in other regions.

However, larger differences are seen when we break the treated pupils down by dosage; that is, by the amount of time they spent using SAM Learning. Students who were not eligible for free school meals were more likely to spend a high amount of time (ten or more hours) using SAM Learning; 28.2% of non-FSM pupils did so in 2018 compared to 20.9% of FSM students. The majority of FSM students who were signed up to SAM Learning did not use it at all; 55.0% in 2018, compared to 45.4% of other students. There were also differences by ethnicity; more than a third (35.5%) of Chinese pupils had a high level of use in 2018, for example, compared with just 17.7% of black pupils. Pupils in some regions tended to make more use of SAM Learning than others. In 2018, the majority (55.3%) of treated pupils who went to a school in the North East had a high level of use, compared to just 16.0% in London, where most treated pupils (59.8%) did not use SAM Learning at all.

4.2 Creating a matched comparison group

We used inverse probability weighting (IPW) to create a comparison group of pupils who were statistically similar to the treated pupils with respect to the confounding variables. This method assign weights to all observations to create a balanced treatment and comparison group. The IPW weights are based on propensity score, as shown below:

$$W_t = 1,\ W_c = \frac{PS}{1 - PS}$$

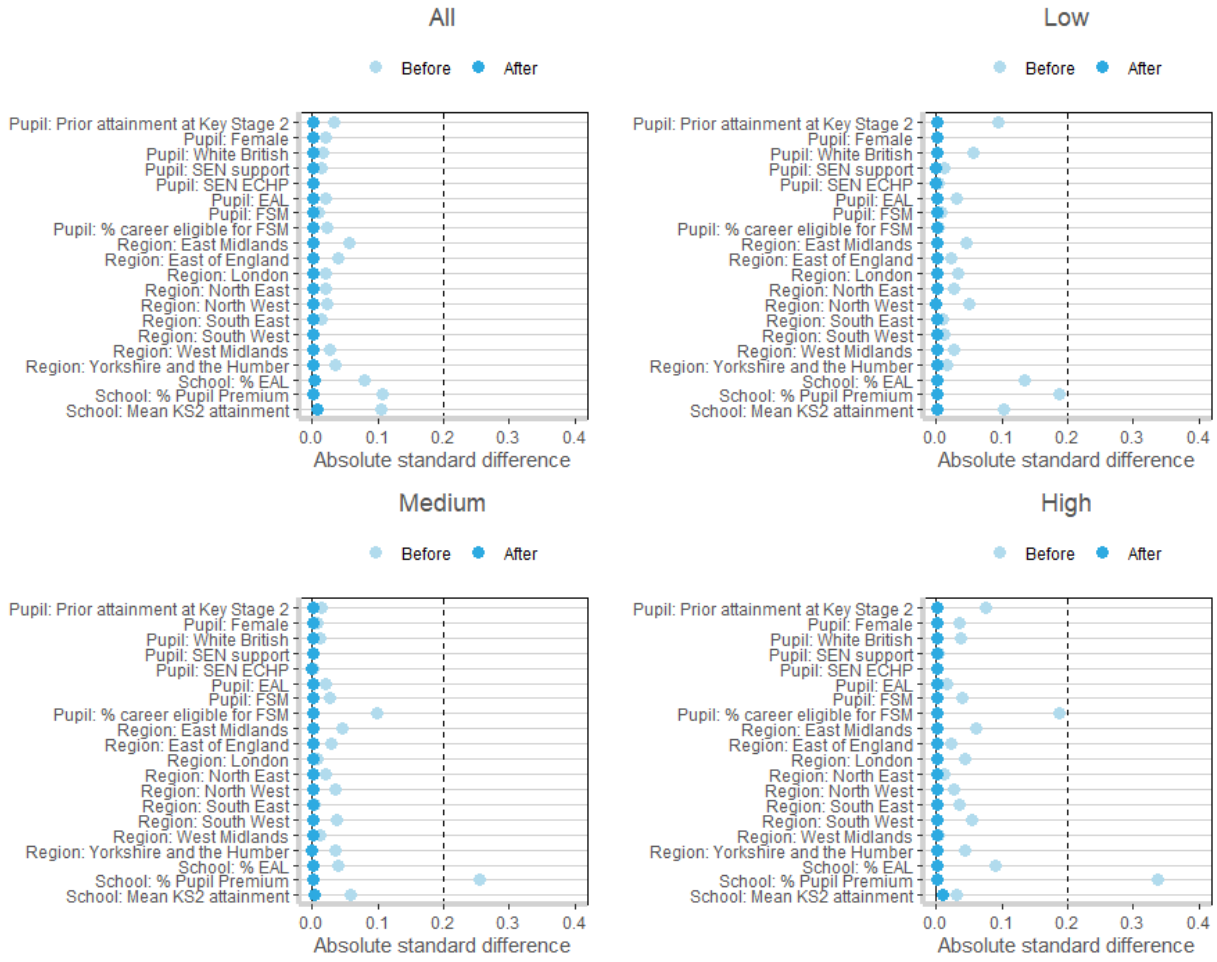
…where Wt is the weight for a treated pupil, Wc is the weight for a comparison pupil, and PS is the propensity score.

The propensity score is a measure the likelihood of each pupil being in the treated group, based on the confounding variables. In this case, the propensity score was calculated by fitting a logistic regression model.

The results obtained from using several alternative methods to create a comparison group are shown in the appendix. IPW was chosen for the main analysis because it created a well-matched comparison group, using the full sample of treated pupils, and was less computationally expensive than some alternative methods.

The graphs shown below in figure 1, known as *loveplots*,[3] show how similar the treated and comparison pupils from the 2018 cohort were to one another, before and after the application of IPW, using a measure called the standardised mean difference. The mean difference is simply the difference between the average value of the variable for the treated students, and the average value for the comparison students. Standardising this measure means that we can compare balance across different variables. Generally, a standardised mean difference of 0.2 or below is considered to indicate good balance. This threshold is shown on the graphs as a dotted line.

Figure 1: Loveplots showing balance before and after IPW weighting, 2018



For the 2018 cohort, the balance was very good. This was the case for the overall group and for the subsets of FSM pupils for every cohort from 2011-18.

[3] Loveplots are named for Professor Thomas E. Love, who first developed them along with colleagues (https://academic.oup.com/eurheartj/article/27/12/1431/647407)

5. Results

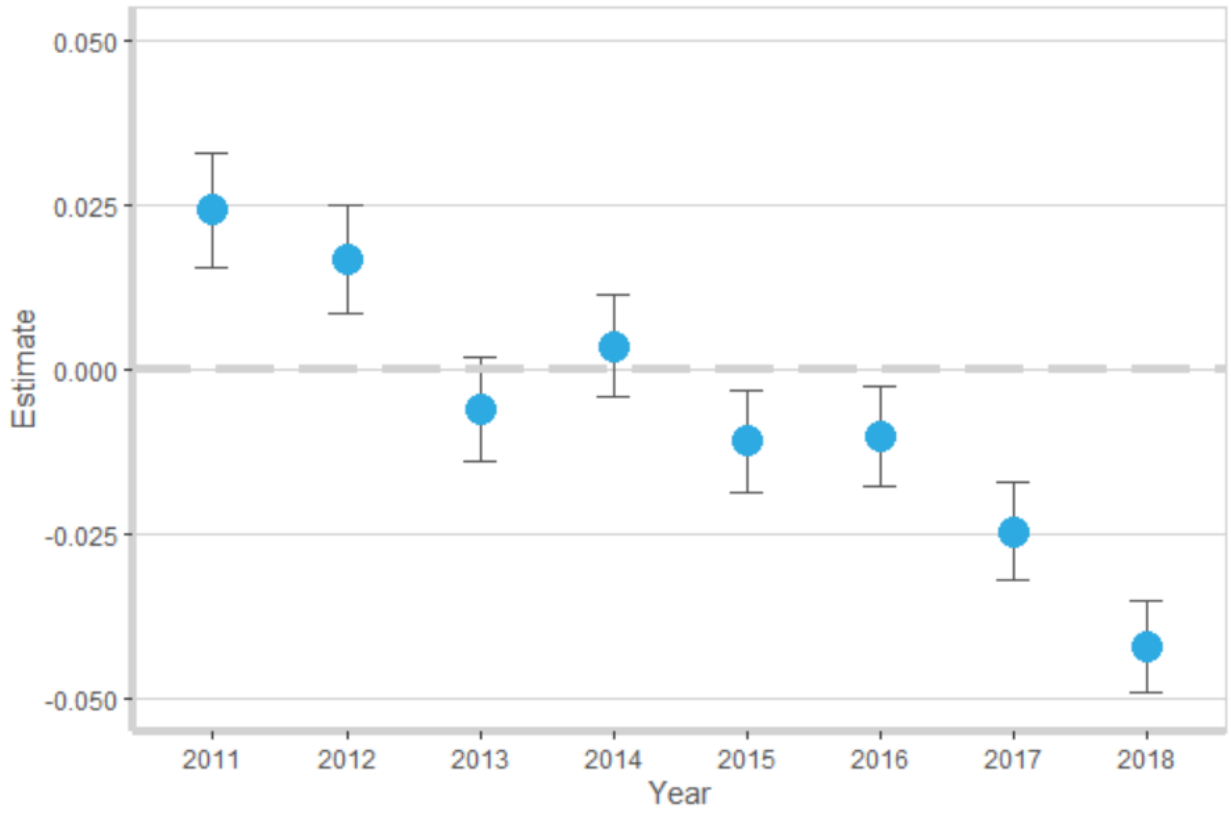
Results are given in the form of both estimated treatment effects and estimated effect sizes. Treatment effects represent the estimated effect of SAM Learning in terms of GCSE grade. GCSE grading changed during the period covered by this report: from 2011-16, six GCSE points was the equivalent of one grade, while from 2017 onwards, one GCSE point was the equivalent of one grade. Effect sizes are a standardised version of the treatment effect; looking at effect sizes allows us to more easily compare the relative size of the effect across all of the outcome years. Where effects sizes are shown, we also include the equivalent months of progress, calculated according to EEF guidelines.[4]

5.1 All pupils

The estimated impact of SAM Learning on all treated pupils is summarised in figure 2. Full results, included estimated impact on GCSE point scores and the equivalent effect sizes, all rounded to two decimal places, are given in table 2. The sample sizes shown in the table give the number of treated pupils included in each model.

We found no consistent significant effect on GCSE point scores for all pupils. The estimated impact is positive in some years, and negative in others.

Figure 2: Estimated effect size on GCSE point scores by year, all pupils



[4] Education Endowment Foundation, 'Evaluation Report Template 2019', accessed from https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/evaluator-resources/writing-a-research-report/

Table 2: Estimated impact on GCSE point scores by year, all pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	0.16	0.25	0.34	-	22,124
Treatment effect	2012	0.08	0.17	0.25	-	24,394
Treatment effect	2013	-0.14	-0.06	0.02	-	26,976
Treatment effect	2014	-0.04	0.04	0.11	-	28,279
Treatment effect	2015	-0.19	-0.11	-0.03	-	30,342
Treatment effect	2016	-0.17	-0.10	-0.03	-	30,926
Treatment effect	2017	-0.06	-0.04	-0.03	-	33,324
Treatment effect	2018	-0.09	-0.08	-0.07	-	37,241
Effect size	2011	0.02	0.02	0.03	0	22,124
Effect size	2012	0.01	0.02	0.02	0	24,394
Effect size	2013	-0.01	-0.01	0.00	0	26,976
Effect size	2014	0.00	0.00	0.01	0	28,279
Effect size	2015	-0.02	-0.01	0.00	0	30,342
Effect size	2016	-0.02	-0.01	0.00	0	30,926
Effect size	2017	-0.03	-0.02	-0.02	0	33,324
Effect size	2018	-0.05	-0.04	-0.04	0	37,241

5.2 Results by dosage

There was considerable variation in how much time pupils spent using SAM Learning. In each cohort, a large proportion, between 30 and 50%, did not use the system at all. Those that did use the system were divided into three dosage groups for the purposes of this evaluation: low use (0-2 hours), medium use (2-10 hours) and high use (10 hours or more). From 2011-2016, pupils were divided fairly evenly between these three groups. However, in 2017 and 2018, a higher proportion of students were in the high dosage group, as shown in table 3.

Table 3: Proportion of SAM Learning pupils in each dosage group, 2011-18

Year	No use	Low	Medium	High
2011	42.5	21.2	18.1	18.2
2012	40.0	23.1	18.0	19.0
2013	45.0	21.3	17.9	15.9
2014	43.4	20.5	19.3	16.8
2015	37.9	22.2	18.8	21.1
2016	40.2	22.1	17.3	20.4
2017	34.7	21.5	14.1	29.7
2018	46.5	16.5	9.7	27.3

5.2.1 High dosage

The estimated impact of SAM Learning on treated pupils in the high dosage group is summarised in figure 3. Full results in given in table 4. The sample sizes shown in the table give the number of treated pupils included in each model.

For the high dosage group, the estimated impact is both positive and significant for every cohort. The effect is the equivalent of between a ninth and a third of a grade per subject.

Figure 3: Estimated effect size on GCSE point scores by year, high dosage

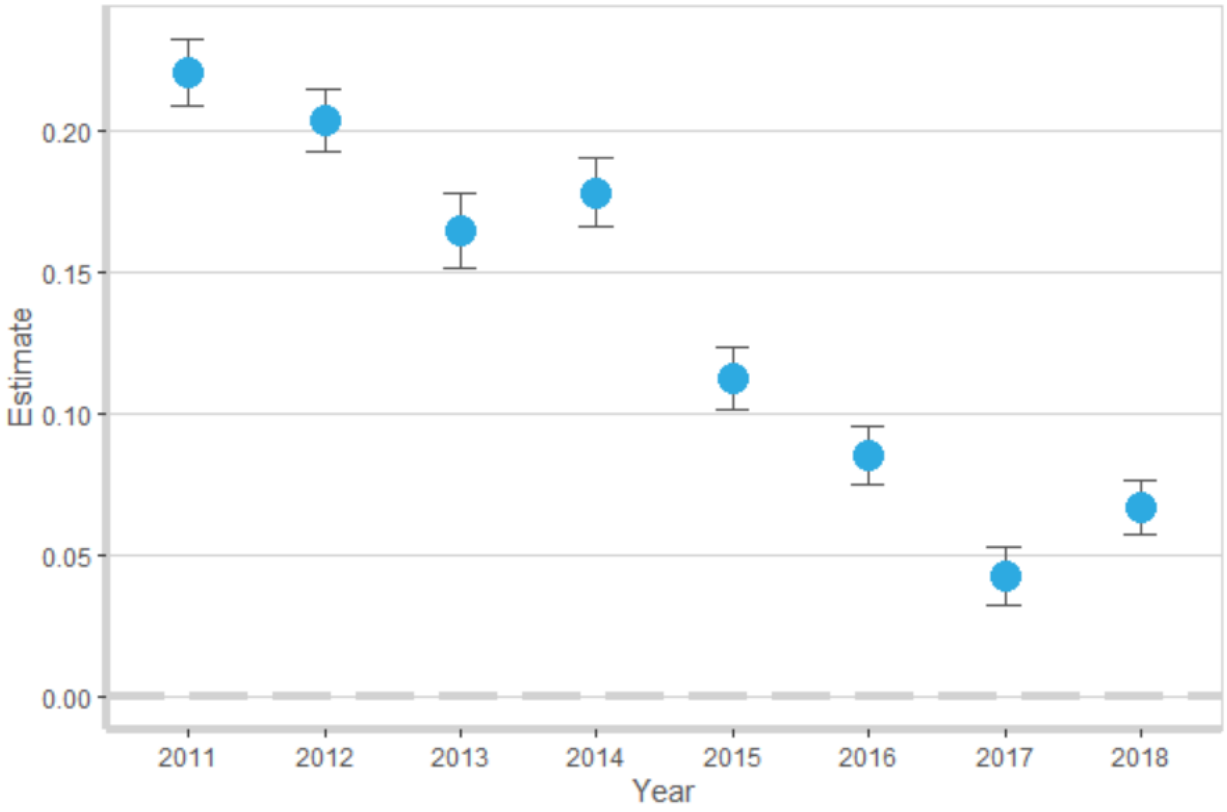


Table 4: Estimated impact on GCSE point scores by year, high dosage pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	2.14	2.26	2.38	-	4,203
Treatment effect	2012	1.93	2.04	2.15	-	4,695
Treatment effect	2013	1.50	1.63	1.76	-	4,343
Treatment effect	2014	1.66	1.78	1.90	-	4,898
Treatment effect	2015	1.02	1.12	1.23	-	6,821
Treatment effect	2016	0.72	0.82	0.92	-	6,563
Treatment effect	2017	0.06	0.08	0.09	-	10,055
Treatment effect	2018	0.11	0.12	0.14	-	10,296
Effect size	2011	0.21	0.22	0.23	3	4,203
Effect size	2012	0.19	0.20	0.21	3	4,695
Effect size	2013	0.15	0.16	0.18	2	4,343
Effect size	2014	0.17	0.18	0.19	2	4,898
Effect size	2015	0.10	0.11	0.12	2	6,821
Effect size	2016	0.08	0.09	0.10	1	6,563
Effect size	2017	0.03	0.04	0.05	0	10,055
Effect size	2018	0.06	0.07	0.08	1	10,296

5.2.2 Medium dosage

The estimated impact of SAM Learning on treated pupils in the medium dosage group is summarised in figure 4. Full results in given in table 5. The sample sizes shown in the table give the number of treated pupils included in each model.

We found no consistent significant effect on GCSE point scores for this group. From 2011-16, we found small but significant positive effects, but small negative effects were found in 2017-18.

Figure 4: Estimated effect size on GCSE point scores by year, medium dosage

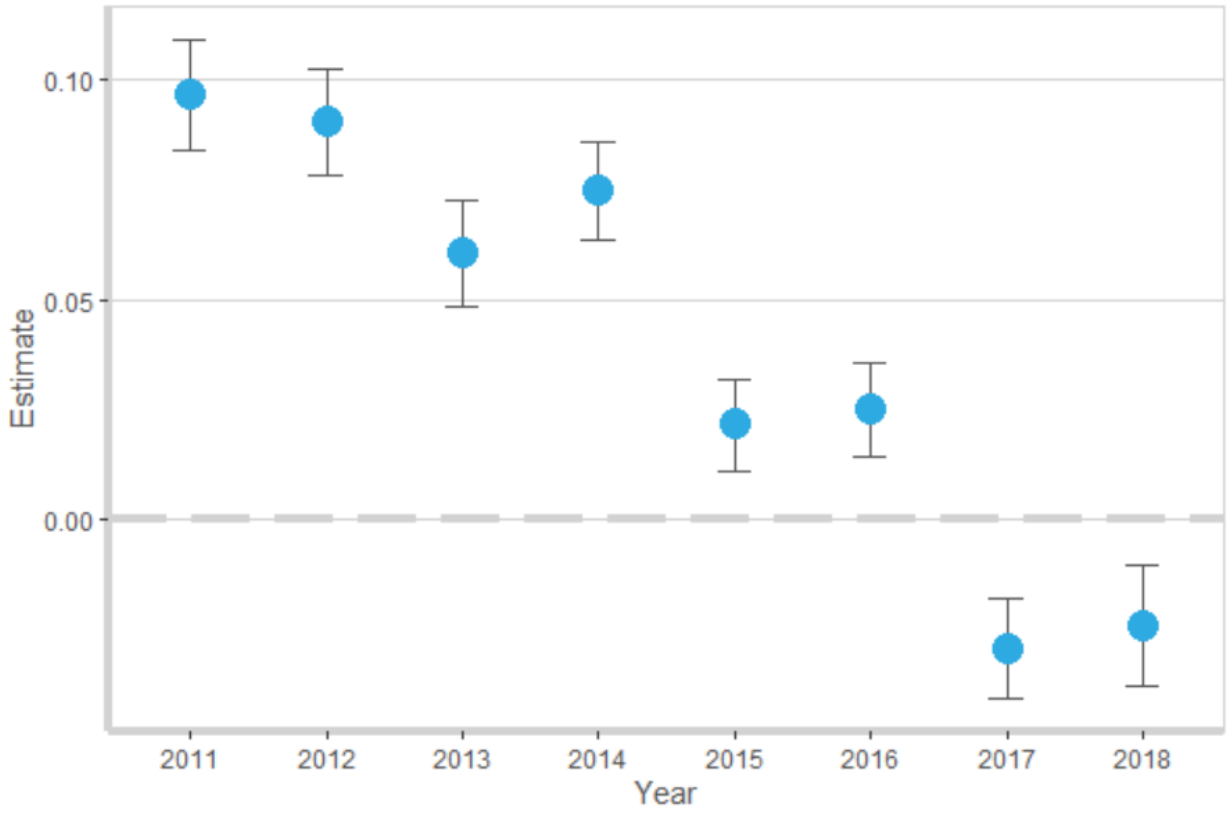


Table 5: Estimated impact on GCSE point scores by year, medium dosage pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	0.86	0.99	1.12	-	4,759
Treatment effect	2012	0.79	0.91	1.03	-	5,662
Treatment effect	2013	0.48	0.60	0.72	-	5,783
Treatment effect	2014	0.63	0.75	0.86	-	5,808
Treatment effect	2015	0.11	0.22	0.32	-	7,048
Treatment effect	2016	0.14	0.24	0.34	-	7,048
Treatment effect	2017	-0.07	-0.05	-0.03	-	7,318
Treatment effect	2018	-0.07	-0.05	-0.02	-	6,272
Effect size	2011	0.08	0.10	0.11	2	4,759
Effect size	2012	0.08	0.09	0.10	1	5,662
Effect size	2013	0.05	0.06	0.07	1	5,783
Effect size	2014	0.06	0.08	0.09	1	5,808

Effect size	2015	0.01	0.02	0.03	0	7,048
Effect size	2016	0.01	0.03	0.04	0	7,048
Effect size	2017	-0.04	-0.03	-0.02	0	7,318
Effect size	2018	-0.04	-0.02	-0.01	0	6,272

5.2.3 Low dosage

The estimated impact of SAM Learning on treated pupils in the low dosage group is summarised in figure 5. Full results in given in table 6. The sample sizes shown in the table give the number of treated pupils included in each model.

We found no consistent significant effect on GCSE point scores for this group. While estimates for some years are positive, in others pupils in this group have significantly lower outcomes than comparison pupils.

Figure 5: Estimated effect size on GCSE point scores by year, low dosage

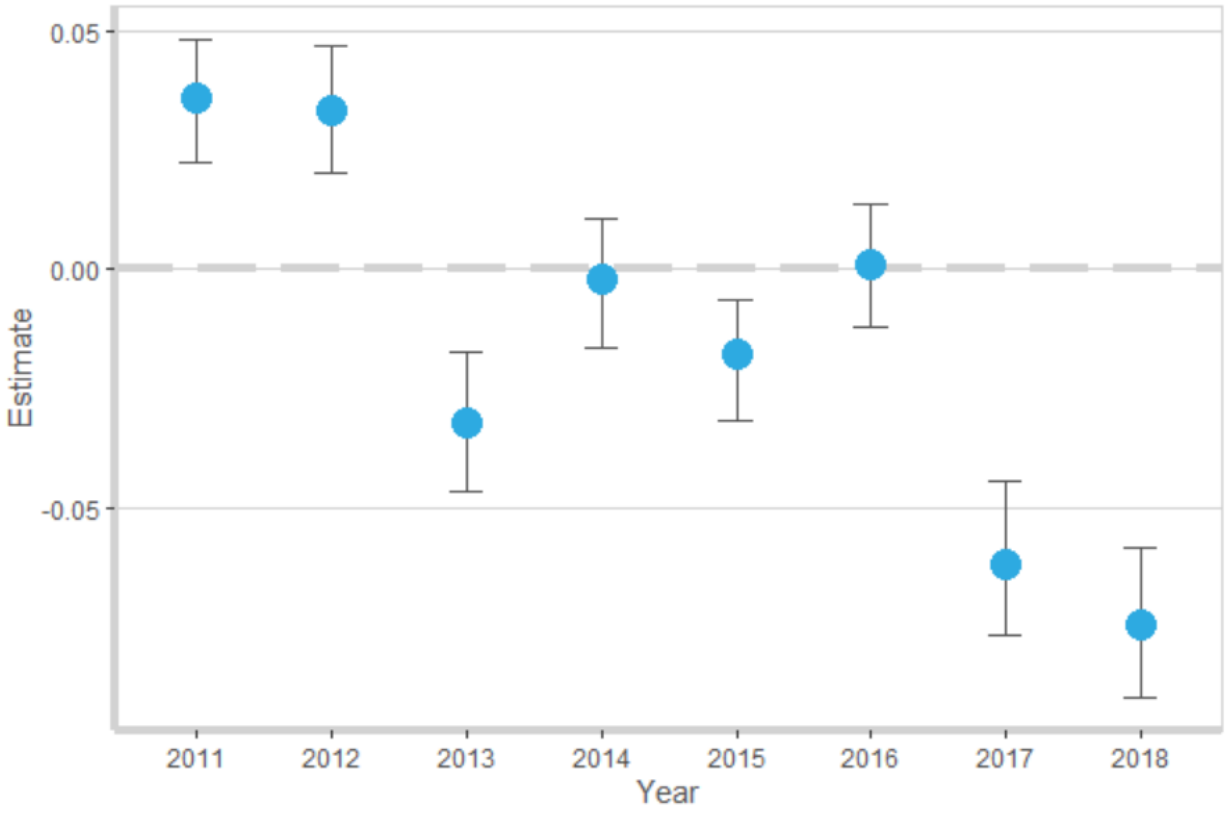


Table 6: Estimated impact on GCSE point scores by year, low dosage pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	0.23	0.37	0.50	-	4,034
Treatment effect	2012	0.20	0.33	0.47	-	4,375
Treatment effect	2013	-0.46	-0.32	-0.17	-	4,902
Treatment effect	2014	-0.16	-0.02	0.11	-	5,391
Treatment effect	2015	-0.31	-0.18	-0.06	-	5,690
Treatment effect	2016	-0.12	0.01	0.13	-	5,506

Treatment effect	2017	-0.14	-0.11	-0.08	-	4,815
Treatment effect	2018	-0.17	-0.14	-0.11	-	3,658
Effect size	2011	0.02	0.04	0.05	0	4,034
Effect size	2012	0.02	0.03	0.05	0	4,375
Effect size	2013	-0.05	-0.03	-0.02	0	4,902
Effect size	2014	-0.02	0.00	0.01	0	5,391
Effect size	2015	-0.03	-0.02	-0.01	0	5,690
Effect size	2016	-0.01	0.00	0.01	0	5,506
Effect size	2017	-0.08	-0.06	-0.04	<0	4,815
Effect size	2018	-0.09	-0.07	-0.06	<0	3,658

5.3 FSM pupils

We also looked at the effects by dosage on pupils who were eligible for free school meals. As discussed in section 3.1, FSM pupils were less likely than other pupils to fall into the high dosage group, and more likely to make no use at all of the product. However, as seen for all students in section 4.2, a higher proportion of FSM pupils fell into the high dosage group in 2017 and 2018. In 2018, for example, 20.9% of FSM students fell into the high dosage group, compared to just 9.9% in the low and 14.1% in the medium.

5.3.1 High dosage

The estimated impact of SAM Learning on treated pupils in the high dosage group who were eligible for free school meals is summarised in figure 6. Full results in given in table 7. The sample sizes shown in the table give the number of treated pupils included in each model.

The estimated impact is both positive and significant for every cohort, the equivalent of between a fifth and more than half a grade per subject. Estimates are lower in more recent years but are still the equivalent of between a fifth and a third of a grade per subject.

Figure 6: Estimated effect size on GCSE point scores by year, high dosage FSM pupils

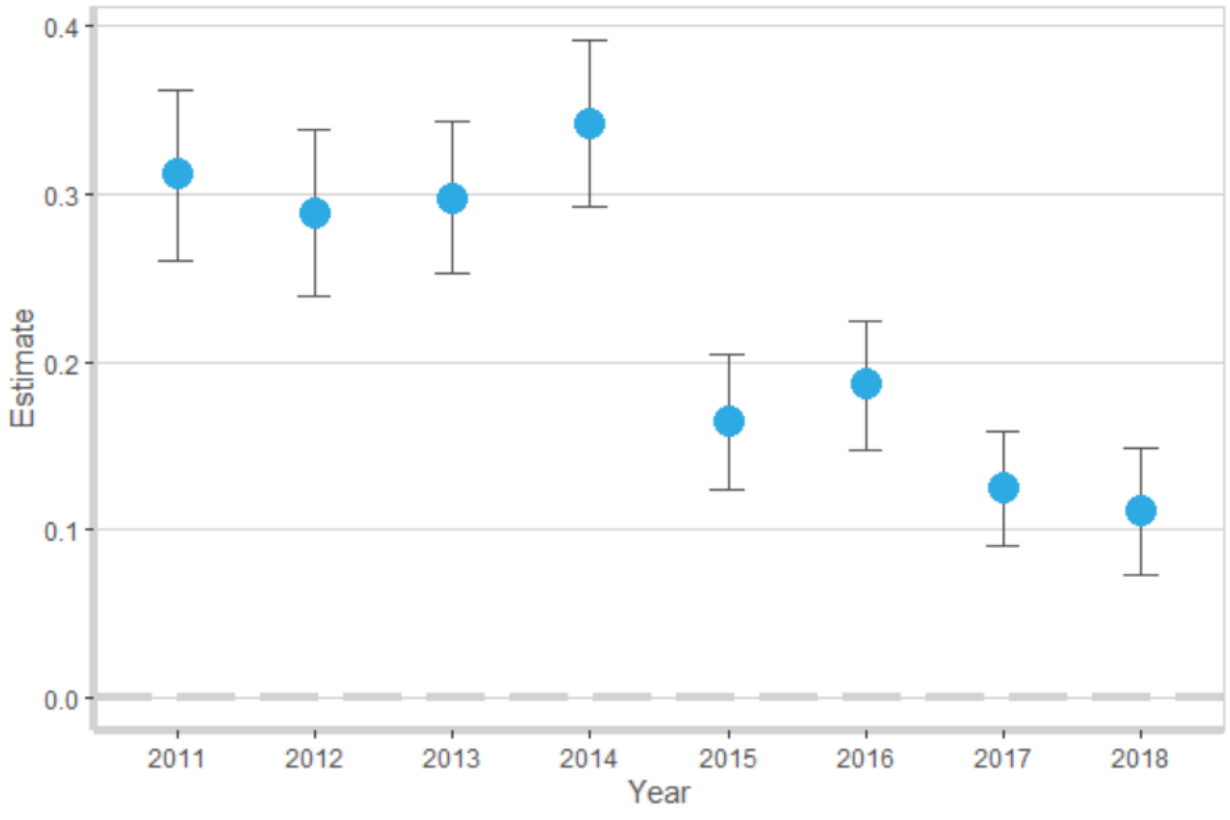


Table 7: Estimated impact on GCSE point scores by year, high dosage FSM pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	2.90	3.47	4.02	-	307
Treatment effect	2012	2.60	3.14	3.69	-	360
Treatment effect	2013	2.73	3.22	3.71	-	358
Treatment effect	2014	3.23	3.78	4.32	-	345
Treatment effect	2015	1.40	1.86	2.31	-	521
Treatment effect	2016	1.56	1.98	2.37	-	526
Treatment effect	2017	0.15	0.21	0.27	-	888
Treatment effect	2018	0.13	0.19	0.26	-	901
Effect size	2011	0.26	0.31	0.36	4	307
Effect size	2012	0.24	0.29	0.34	4	360
Effect size	2013	0.25	0.30	0.34	4	358
Effect size	2014	0.29	0.34	0.39	4	345
Effect size	2015	0.12	0.17	0.21	2	521
Effect size	2016	0.15	0.19	0.22	3	526
Effect size	2017	0.09	0.12	0.16	2	888
Effect size	2018	0.07	0.11	0.15	2	901

5.3.2 Medium dosage

The estimated impact of SAM Learning on treated pupils in the medium dosage group who were eligible for free school meals is summarised in figure 7. Full results in given in table 8. The sample sizes shown in the table give the number of treated pupils included in each model.

Estimates for this group were positive and significant for each year from 2011-6, but were inconclusive in 2017 and 2018.

Figure 7: Estimated effect size on GCSE point scores by year, medium dosage FSM pupils

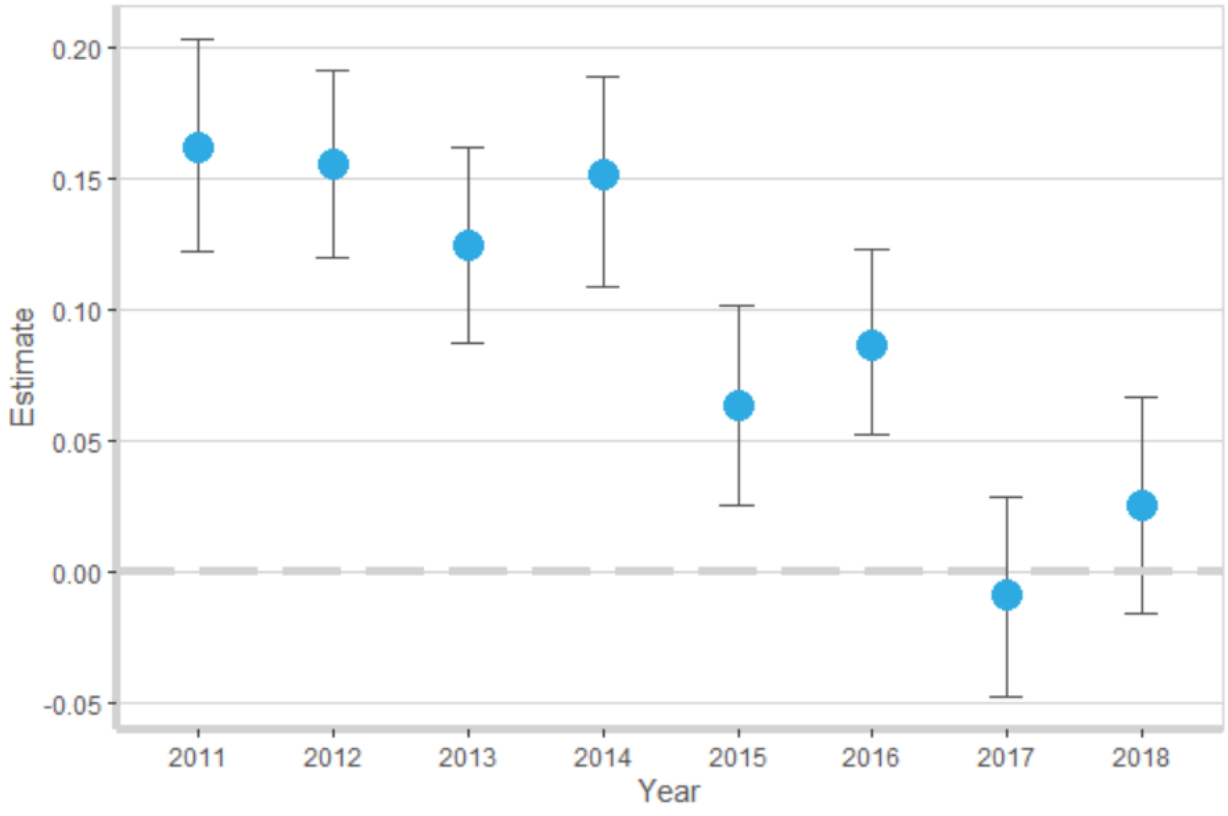


Table 8: Estimated impact on GCSE point scores by year, medium dosage FSM pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	1.36	1.80	2.26	-	508
Treatment effect	2012	1.31	1.69	2.08	-	656
Treatment effect	2013	0.94	1.35	1.74	-	650
Treatment effect	2014	1.20	1.67	2.08	-	558
Treatment effect	2015	0.29	0.72	1.14	-	749
Treatment effect	2016	0.55	0.91	1.29	-	706
Treatment effect	2017	-0.08	-0.02	0.05	-	817
Treatment effect	2018	-0.03	0.04	0.12	-	623
Effect size	2011	0.12	0.16	0.20	2	508
Effect size	2012	0.12	0.16	0.19	2	656
Effect size	2013	0.09	0.12	0.16	2	650
Effect size	2014	0.11	0.15	0.19	2	558
Effect size	2015	0.03	0.06	0.10	1	749
Effect size	2016	0.05	0.09	0.12	1	706
Effect size	2017	-0.05	-0.01	0.03	0	817
Effect size	2018	-0.02	0.03	0.07	0	623

5.3.3 Low dosage

The estimated impact of SAM Learning on treated pupils in the low dosage group who were eligible for free school meals is summarised in figure 8. Full results in given in table 9. The sample sizes shown in the table give the number of treated pupils included in each model.

Results for this group were inconclusive; while there were significant positive estimates for some of the outcome years, estimates for other years were insignificant or negative.

Figure 8: Estimated effect size on GCSE point scores by year, low dosage FSM pupils

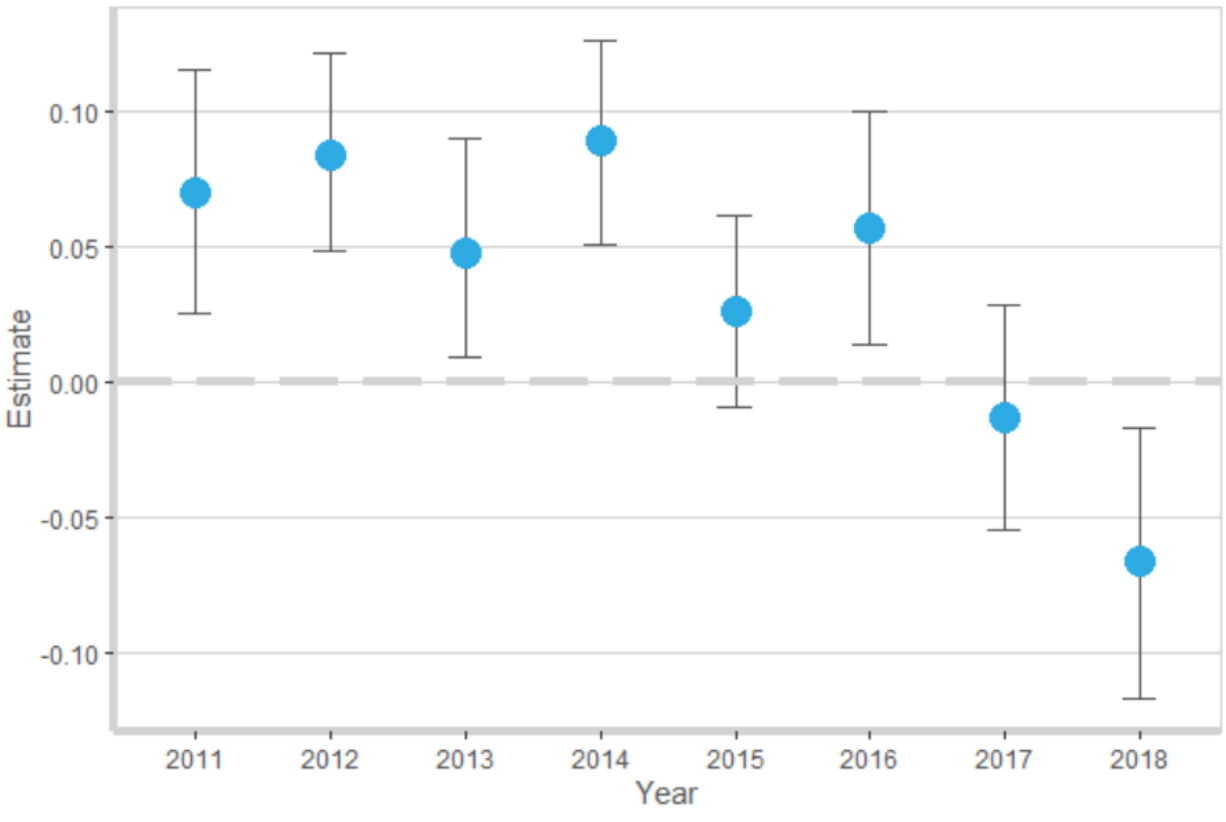


Table 9: Estimated impact on GCSE point scores by year, low dosage FSM pupils

Type	Year	Lower CI	Estimate	Upper CI	Months progress	Sample size
Treatment effect	2011	0.28	0.78	1.29	-	520
Treatment effect	2012	0.53	0.92	1.32	-	666
Treatment effect	2013	0.10	0.52	0.98	-	602
Treatment effect	2014	0.56	0.99	1.40	-	714
Treatment effect	2015	-0.10	0.30	0.70	-	727
Treatment effect	2016	0.15	0.61	1.05	-	664
Treatment effect	2017	-0.09	-0.02	0.05	-	609
Treatment effect	2018	-0.20	-0.11	-0.03	-	433
Effect size	2011	0.03	0.07	0.12	1	520
Effect size	2012	0.05	0.08	0.12	1	666
Effect size	2013	0.01	0.05	0.09	1	602
Effect size	2014	0.05	0.09	0.13	1	714
Effect size	2015	-0.01	0.03	0.06	0	727
Effect size	2016	0.01	0.06	0.10	1	664
Effect size	2017	-0.05	-0.01	0.03	0	609
Effect size	2018	-0.12	-0.07	-0.02	<0	433

6. Discussion

6.1 Overview

In terms of evaluating the impact of SAM Learning on GCSE grades, we found significant positive effects on high dosage users, of between a ninth and a third of a grade per subject. Effects were stronger for high dosage pupils who were eligible for free school meals, at between a fifth and more than half a grade per subject. However, we found no conclusive evidence of an overall effect from having a SAM Learning account, including pupils at all dosage levels and those with zero use.

6.2 Limitations

Many of the limitations associated with this evaluation arise from the fact that treated and comparison pupils were matched using observational data from the National Pupil Database (NPD). The NPD is, of course, limited in scope. 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.

In this case, the omission of historical school-level GCSE attainment from the matching variables, as discussed in section 3.1, could have resulted in bias if treated pupils were matched to comparison pupils at better or worse performing schools. However, we judge that including these variables would have been even more likely to introduce bias, considering the issues with reliability of the data and the long-term nature of many schools' involvement with SAM Learning. There may also be differences in pupil motivation that we were unable to account for; for example, pupils in the low dosage group may be those with low motivation, resulting in underestimates of SAM Learning's impact if they were matched with otherwise similar pupils with higher levels of motivation. It is also possible that the significant positive effects found for the high dosage group are at least partly due to unobserved differences. This is likely to be an issue for most evaluations looking at different dosage levels using a quasi-experimental design; any intervention that allows the user some choice on their level of engagement will, almost by definition, end up with more motivated pupils in the higher dosage groups. See the appendix for more discussion on these points and the presentation of results obtained using alternative methodology.

Some comparison pupils may have used similar products. If this was the case, our analysis would not be an evaluation of the SAM Learning 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 SAM Learning, assuming that the equivalent support had a positive effect on some comparison pupils' outcomes. We would note, however, that not controlling for this effect may be the relevant analysis as it represents an evaluation of SAM Learning against current conditions, with schools' and / or pupils' choices to engage with other projects being included in the makeup of controls.

Finally, the nature of the product means that a high proportion of 'treated' pupils did not actually engage with SAM Learning at all. When a school subscribes to SAM Learning, all pupils at the school are given an account. In some cases, schools do not choose to make use of all their pupil accounts. In other cases, schools may have encouraged pupils to make use of their accounts but pupils have failed to do so. Ideally, we would exclude the former cases from the treated group, as there was no intention to treat these pupils, but include the latter. However, there is no way of distinguishing between the two from the available dataset.

7. Appendix: Sensitivity analysis

In this section, we present the results obtained from analysis completed using alternative methodology. Comparing the results from these alternative approaches to those in our main analysis can give an indication as to how robust the results are.

7.1 Alternative methods for mitigating confounding effects

This section shows the results obtained using three alternative methods for mitigating confounding effects. The three methods are: nearest neighbor matching based on propensity scores (NN), coarsened exact matching (CEM) and weighting using covariate balancing propensity scores (CBPS). Estimates obtained using CBPS did not use bootstrapping, so the confidence intervals do not account for the uncertainty from the matching process and may be artificially narrow.

Comparing the estimates produced by the different matching methods, we can see that the NN method had a slight tendency to overestimate compared to the other two. However, CEM and CBPS tended to agree with the IPW estimates shown in the body of the report, suggesting that these results are fairly robust to the use of different matching methods.

7.1.1 Estimated impact on GCSE point score

The table below includes the estimated impact of SAM Learning on average uncapped GCSE point scores. GCSE grading changed during the period covered by this report: from 2011-16, six GCSE points was the equivalent of one grade, while from 2017 onwards, one GCSE point was the equivalent of one grade. Post-reform years are shown in italics to emphasise this point.

Method	Dosage	FSM status	Cohort	Lower CI	Estimate	Upper CI
CEM	All	All	2011	0.17	0.28	0.38
CEM	All	All	2012	0.15	0.25	0.35
CEM	All	All	2013	-0.18	-0.09	0.00
CEM	All	All	2014	-0.07	0.01	0.10
CEM	All	All	2015	-0.14	-0.05	0.03
CEM	All	All	2016	-0.15	-0.07	0.00
CEM	*All*	*All*	*2017*	*-0.04*	*-0.02*	*0.00*
CEM	*All*	*All*	*2018*	*-0.07*	*-0.05*	*-0.04*
CEM	Low	All	2011	0.05	0.24	0.43
CEM	Low	All	2012	0.03	0.23	0.45
CEM	Low	All	2013	-0.72	-0.52	-0.33
CEM	Low	All	2014	-0.38	-0.22	-0.05
CEM	Low	All	2015	-0.55	-0.38	-0.21
CEM	Low	All	2016	-0.34	-0.16	0.00
CEM	*Low*	*All*	*2017*	*-0.10*	*-0.06*	*-0.02*
CEM	*Low*	*All*	*2018*	*-0.14*	*-0.10*	*-0.06*
CEM	Med	All	2011	0.65	0.81	0.98
CEM	Med	All	2012	0.47	0.66	0.85
CEM	Med	All	2013	0.58	0.74	0.92
CEM	Med	All	2014	0.43	0.61	0.79

CEM	Med	All	2015	0.00	0.16	0.32
CEM	Med	All	2016	-0.02	0.13	0.27
CEM	*Med*	*All*	*2017*	*-0.07*	*-0.04*	*-0.01*
CEM	*Med*	*All*	*2018*	*-0.07*	*-0.04*	*0.00*
CEM	High	All	2011	1.69	1.87	2.06
CEM	High	All	2012	1.76	1.93	2.11
CEM	High	All	2013	1.35	1.53	1.70
CEM	High	All	2014	1.56	1.73	1.90
CEM	High	All	2015	0.98	1.13	1.29
CEM	High	All	2016	0.72	0.87	1.00
CEM	*High*	*All*	*2017*	*0.10*	*0.12*	*0.15*
CEM	*High*	*All*	*2018*	*0.09*	*0.12*	*0.15*
CEM	Low	FSM	2011	0.65	1.54	2.44
CEM	Low	FSM	2012	-0.61	0.19	0.98
CEM	Low	FSM	2013	-0.91	-0.21	0.45
CEM	Low	FSM	2014	-0.09	0.56	1.23
CEM	Low	FSM	2015	0.02	0.80	1.55
CEM	Low	FSM	2016	0.36	1.11	1.86
CEM	*Low*	*FSM*	*2017*	*-0.07*	*0.05*	*0.16*
CEM	*Low*	*FSM*	*2018*	*-0.29*	*-0.04*	*0.00*
CEM	Med	FSM	2011	1.96	2.79	3.55
CEM	Med	FSM	2012	1.18	1.87	2.52
CEM	Med	FSM	2013	1.64	2.34	3.06
CEM	Med	FSM	2014	0.43	1.08	1.72
CEM	Med	FSM	2015	-0.09	0.56	1.17
CEM	Med	FSM	2016	0.86	1.43	2.01
CEM	*Med*	*FSM*	*2017*	*-0.09*	*-0.01*	*0.07*
CEM	*Med*	*FSM*	*2018*	*0.01*	*0.12*	*0.22*
CEM	High	FSM	2011	2.45	3.30	4.14
CEM	High	FSM	2012	2.37	3.14	3.90
CEM	High	FSM	2013	2.41	3.27	4.17
CEM	High	FSM	2014	2.58	3.47	4.28
CEM	High	FSM	2015	1.11	1.85	2.55
CEM	High	FSM	2016	1.58	2.28	2.93
CEM	*High*	*FSM*	*2017*	*0.11*	*0.20*	*0.29*
CEM	*High*	*FSM*	*2018*	*0.19*	*0.28*	*0.37*
NN	All	All	2011	0.23	0.34	0.45
NN	All	All	2012	0.33	0.44	0.53
NN	All	All	2013	0.05	0.14	0.24
NN	All	All	2014	0.19	0.28	0.37
NN	All	All	2015	0.06	0.15	0.25
NN	All	All	2016	-0.02	0.07	0.16
NN	*All*	*All*	*2017*	*-0.02*	*-0.01*	*0.01*
NN	*All*	*All*	*2018*	*-0.06*	*-0.04*	*-0.03*
NN	Low	All	2011	0.22	0.44	0.66

NN	Low	All	2012	0.43	0.65	0.86
NN	Low	All	2013	-0.24	-0.03	0.19
NN	Low	All	2014	0.12	0.31	0.50
NN	Low	All	2015	-0.09	0.10	0.30
NN	Low	All	2016	0.03	0.22	0.41
NN	*Low*	*All*	*2017*	*-0.12*	*-0.08*	*-0.04*
NN	*Low*	*All*	*2018*	*-0.12*	*-0.07*	*-0.02*
NN	Med	All	2011	1.28	1.47	1.67
NN	Med	All	2012	1.02	1.21	1.40
NN	Med	All	2013	0.77	0.95	1.13
NN	Med	All	2014	0.93	1.12	1.32
NN	Med	All	2015	0.37	0.54	0.70
NN	Med	All	2016	0.36	0.51	0.67
NN	*Med*	*All*	*2017*	*0.01*	*0.04*	*0.07*
NN	*Med*	*All*	*2018*	*-0.01*	*0.02*	*0.06*
NN	High	All	2011	2.47	2.67	2.86
NN	High	All	2012	1.95	2.13	2.31
NN	High	All	2013	1.79	2.00	2.21
NN	High	All	2014	1.94	2.12	2.30
NN	High	All	2015	1.25	1.42	1.58
NN	High	All	2016	0.86	1.02	1.18
NN	*High*	*All*	*2017*	*0.08*	*0.11*	*0.13*
NN	*High*	*All*	*2018*	*0.15*	*0.17*	*0.20*
NN	Low	FSM	2011	-0.23	0.53	1.31
NN	Low	FSM	2012	0.69	1.31	1.93
NN	Low	FSM	2013	-0.55	0.07	0.72
NN	Low	FSM	2014	0.12	0.74	1.37
NN	Low	FSM	2015	-0.29	0.32	0.95
NN	Low	FSM	2016	-0.11	0.60	1.30
NN	*Low*	*FSM*	*2017*	*-0.18*	*-0.07*	*0.04*
NN	*Low*	*FSM*	*2018*	*-0.27*	*-0.12*	*0.02*
NN	Med	FSM	2011	0.85	1.58	2.30
NN	Med	FSM	2012	0.74	1.39	2.03
NN	Med	FSM	2013	0.59	1.19	1.79
NN	Med	FSM	2014	0.91	1.58	2.26
NN	Med	FSM	2015	0.39	1.04	1.72
NN	Med	FSM	2016	0.40	1.01	1.63
NN	*Med*	*FSM*	*2017*	*-0.18*	*-0.08*	*0.02*
NN	*Med*	*FSM*	*2018*	*-0.08*	*0.02*	*0.14*
NN	High	FSM	2011	2.12	2.99	3.86
NN	High	FSM	2012	2.94	3.76	4.58
NN	High	FSM	2013	2.13	2.90	3.68
NN	High	FSM	2014	3.66	4.57	5.46
NN	High	FSM	2015	0.80	1.52	2.25
NN	High	FSM	2016	0.84	1.55	2.27

NN	*High*	*FSM*	*2017*	*0.12*	*0.20*	*0.29*
NN	*High*	*FSM*	*2018*	*0.13*	*0.22*	*0.31*
CBPS	All	All	2011	0.23	0.27	0.31
CBPS	All	All	2012	0.15	0.19	0.22
CBPS	All	All	2013	-0.12	-0.08	-0.05
CBPS	All	All	2014	-0.02	0.01	0.05
CBPS	All	All	2015	-0.18	-0.15	-0.11
CBPS	All	All	2016	-0.19	-0.15	-0.12
CBPS	*All*	*All*	*2017*	*-0.06*	*-0.05*	*-0.05*
CBPS	*All*	*All*	*2018*	*-0.09*	*-0.08*	*-0.08*
CBPS	Low	All	2011	0.32	0.35	0.39
CBPS	Low	All	2012	0.32	0.36	0.39
CBPS	Low	All	2013	-0.41	-0.38	-0.34
CBPS	Low	All	2014	-0.13	-0.09	-0.06
CBPS	Low	All	2015	-0.27	-0.24	-0.20
CBPS	Low	All	2016	-0.10	-0.07	-0.03
CBPS	*Low*	*All*	*2017*	*-0.12*	*-0.12*	*-0.11*
CBPS	*Low*	*All*	*2018*	*-0.15*	*-0.14*	*-0.13*
CBPS	Med	All	2011	0.98	1.01	1.04
CBPS	Med	All	2012	0.89	0.92	0.95
CBPS	Med	All	2013	0.52	0.55	0.59
CBPS	Med	All	2014	0.65	0.69	0.72
CBPS	Med	All	2015	0.15	0.18	0.22
CBPS	Med	All	2016	0.14	0.17	0.20
CBPS	*Med*	*All*	*2017*	*-0.07*	*-0.06*	*-0.06*
CBPS	*Med*	*All*	*2018*	*-0.06*	*-0.05*	*-0.05*
CBPS	High	All	2011	2.26	2.29	2.32
CBPS	High	All	2012	2.00	2.03	2.06
CBPS	High	All	2013	1.58	1.61	1.64
CBPS	High	All	2014	1.68	1.71	1.74
CBPS	High	All	2015	1.04	1.07	1.11
CBPS	High	All	2016	0.70	0.74	0.77
CBPS	*High*	*All*	*2017*	*0.06*	*0.06*	*0.07*
CBPS	*High*	*All*	*2018*	*0.11*	*0.12*	*0.13*

7.1.2 Estimated effect size

Method	Dosage	FSM status	Cohort	Lower CI	Estimate	Upper CI
CEM	All	All	2011	0.02	0.03	0.04
CEM	All	All	2012	0.01	0.03	0.04
CEM	All	All	2013	-0.02	-0.01	0.00
CEM	All	All	2014	-0.01	0.00	0.01
CEM	All	All	2015	-0.01	-0.01	0.00
CEM	All	All	2016	-0.02	-0.01	0.00

CEM	All	All	2017	-0.02	-0.01	0.00
CEM	All	All	2018	-0.04	-0.03	-0.02
CEM	Low	All	2011	0.01	0.02	0.04
CEM	Low	All	2012	0.00	0.02	0.05
CEM	Low	All	2013	-0.07	-0.05	-0.03
CEM	Low	All	2014	-0.04	-0.02	0.00
CEM	Low	All	2015	-0.06	-0.04	-0.02
CEM	Low	All	2016	-0.04	-0.02	0.00
CEM	Low	All	2017	-0.06	-0.03	-0.01
CEM	Low	All	2018	-0.08	-0.05	-0.03
CEM	Med	All	2011	0.06	0.08	0.10
CEM	Med	All	2012	0.05	0.07	0.09
CEM	Med	All	2013	0.06	0.08	0.09
CEM	Med	All	2014	0.04	0.06	0.08
CEM	Med	All	2015	0.00	0.02	0.03
CEM	Med	All	2016	0.00	0.01	0.03
CEM	Med	All	2017	-0.04	-0.02	0.00
CEM	Med	All	2018	-0.04	-0.02	0.00
CEM	High	All	2011	0.17	0.18	0.20
CEM	High	All	2012	0.18	0.19	0.21
CEM	High	All	2013	0.14	0.15	0.17
CEM	High	All	2014	0.16	0.17	0.19
CEM	High	All	2015	0.10	0.11	0.13
CEM	High	All	2016	0.07	0.09	0.10
CEM	High	All	2017	0.05	0.07	0.08
CEM	High	All	2018	0.05	0.07	0.08
CEM	Low	FSM	2011	0.06	0.14	0.22
CEM	Low	FSM	2012	-0.06	0.02	0.09
CEM	Low	FSM	2013	-0.08	-0.02	0.04
CEM	Low	FSM	2014	-0.01	0.05	0.11
CEM	Low	FSM	2015	0.00	0.07	0.14
CEM	Low	FSM	2016	0.03	0.11	0.18
CEM	Low	FSM	2017	-0.04	0.03	0.09
CEM	Low	FSM	2018	-0.16	-0.02	0.00
CEM	Med	FSM	2011	0.18	0.25	0.32
CEM	Med	FSM	2012	0.11	0.17	0.23
CEM	Med	FSM	2013	0.15	0.22	0.28
CEM	Med	FSM	2014	0.04	0.10	0.16
CEM	Med	FSM	2015	-0.01	0.05	0.10
CEM	Med	FSM	2016	0.08	0.14	0.19
CEM	Med	FSM	2017	-0.05	-0.01	0.04
CEM	Med	FSM	2018	0.01	0.07	0.13
CEM	High	FSM	2011	0.22	0.30	0.37
CEM	High	FSM	2012	0.22	0.29	0.36
CEM	High	FSM	2013	0.22	0.30	0.39

CEM	High	FSM	2014	0.23	0.31	0.39
CEM	High	FSM	2015	0.10	0.16	0.23
CEM	High	FSM	2016	0.15	0.22	0.28
CEM	High	FSM	2017	0.07	0.12	0.17
CEM	High	FSM	2018	0.11	0.16	0.21
NN	All	All	2011	0.02	0.03	0.04
NN	All	All	2012	0.03	0.04	0.05
NN	All	All	2013	0.00	0.01	0.02
NN	All	All	2014	0.02	0.03	0.04
NN	All	All	2015	0.01	0.02	0.02
NN	All	All	2016	0.00	0.01	0.02
NN	All	All	2017	-0.01	0.00	0.00
NN	All	All	2018	-0.03	-0.02	-0.01
NN	Low	All	2011	0.02	0.04	0.06
NN	Low	All	2012	0.04	0.06	0.09
NN	Low	All	2013	-0.02	0.00	0.02
NN	Low	All	2014	0.01	0.03	0.05
NN	Low	All	2015	-0.01	0.01	0.03
NN	Low	All	2016	0.00	0.02	0.04
NN	Low	All	2017	-0.07	-0.04	-0.02
NN	Low	All	2018	-0.06	-0.04	-0.01
NN	Med	All	2011	0.12	0.14	0.16
NN	Med	All	2012	0.10	0.12	0.14
NN	Med	All	2013	0.08	0.10	0.11
NN	Med	All	2014	0.09	0.11	0.13
NN	Med	All	2015	0.04	0.05	0.07
NN	Med	All	2016	0.04	0.05	0.07
NN	Med	All	2017	0.00	0.02	0.04
NN	Med	All	2018	-0.01	0.01	0.03
NN	High	All	2011	0.24	0.26	0.28
NN	High	All	2012	0.19	0.21	0.23
NN	High	All	2013	0.18	0.20	0.22
NN	High	All	2014	0.19	0.21	0.23
NN	High	All	2015	0.12	0.14	0.16
NN	High	All	2016	0.09	0.11	0.12
NN	High	All	2017	0.05	0.06	0.07
NN	High	All	2018	0.08	0.09	0.11
NN	Low	FSM	2011	-0.02	0.05	0.12
NN	Low	FSM	2012	0.06	0.12	0.18
NN	Low	FSM	2013	-0.05	0.01	0.07
NN	Low	FSM	2014	0.01	0.07	0.12
NN	Low	FSM	2015	-0.03	0.03	0.08
NN	Low	FSM	2016	-0.01	0.06	0.12
NN	Low	FSM	2017	-0.11	-0.04	0.02
NN	Low	FSM	2018	-0.15	-0.07	0.01

NN	Med	FSM	2011	0.08	0.14	0.21
NN	Med	FSM	2012	0.07	0.13	0.19
NN	Med	FSM	2013	0.05	0.11	0.17
NN	Med	FSM	2014	0.08	0.14	0.20
NN	Med	FSM	2015	0.03	0.09	0.15
NN	Med	FSM	2016	0.04	0.10	0.15
NN	Med	FSM	2017	-0.11	-0.05	0.01
NN	Med	FSM	2018	-0.05	0.01	0.08
NN	High	FSM	2011	0.19	0.27	0.35
NN	High	FSM	2012	0.27	0.35	0.42
NN	High	FSM	2013	0.20	0.27	0.34
NN	High	FSM	2014	0.33	0.41	0.50
NN	High	FSM	2015	0.07	0.14	0.20
NN	High	FSM	2016	0.08	0.15	0.22
NN	High	FSM	2017	0.07	0.12	0.17
NN	High	FSM	2018	0.07	0.13	0.18
CBPS	All	All	2011	0.02	0.03	0.03
CBPS	All	All	2012	0.02	0.02	0.02
CBPS	All	All	2013	-0.01	-0.01	0.00
CBPS	All	All	2014	0.00	0.00	0.01
CBPS	All	All	2015	-0.02	-0.01	-0.01
CBPS	All	All	2016	-0.02	-0.02	-0.01
CBPS	All	All	2017	-0.03	-0.03	-0.03
CBPS	All	All	2018	-0.05	-0.05	-0.04
CBPS	Low	All	2011	0.03	0.04	0.04
CBPS	Low	All	2012	0.03	0.04	0.04
CBPS	Low	All	2013	-0.04	-0.04	-0.04
CBPS	Low	All	2014	-0.01	-0.01	-0.01
CBPS	Low	All	2015	-0.03	-0.02	-0.02
CBPS	Low	All	2016	-0.01	-0.01	0.00
CBPS	Low	All	2017	-0.07	-0.06	-0.06
CBPS	Low	All	2018	-0.08	-0.08	-0.07
CBPS	Med	All	2011	0.10	0.10	0.10
CBPS	Med	All	2012	0.09	0.09	0.10
CBPS	Med	All	2013	0.05	0.06	0.06
CBPS	Med	All	2014	0.07	0.07	0.07
CBPS	Med	All	2015	0.02	0.02	0.02
CBPS	Med	All	2016	0.01	0.02	0.02
CBPS	Med	All	2017	-0.04	-0.03	-0.03
CBPS	Med	All	2018	-0.03	-0.03	-0.02
CBPS	High	All	2011	0.23	0.23	0.23
CBPS	High	All	2012	0.20	0.21	0.21
CBPS	High	All	2013	0.16	0.17	0.17
CBPS	High	All	2014	0.17	0.17	0.18
CBPS	High	All	2015	0.11	0.11	0.11

CBPS	High	All	2016	0.07	0.08	0.08
CBPS	High	All	2017	0.03	0.04	0.04
CBPS	High	All	2018	0.06	0.06	0.07

7.2 Alternative groups of pupils

In this section, we explore the impact of SAM Learning on two subgroups of pupils: those pupils who attended a school that had recently joined SAM Learning, and *zero dosage* pupils; that is, pupils who had a SAM Learning account but did not use it.

7.2.1 Recent joiners

7.2.1.1 Performance of recent joiners

In our main analysis, we found that estimates of the impact of SAM Learning were lower in more recent years. From this analysis, it wasn't clear why this might be; it is possible that SAM Learning made changes to their platform in recent years, or that the product was less effective following recent changes to the education system. Alternatively, it could be that SAM Learning is more effective in schools that have recently signed up, and the lower estimates in recent years reflect the higher proportion of long term users in the sample. To explore this, we looked at the impact of SAM Learning on outcomes in 2017/18 for those pupils in schools that joined SAM Learning in 2016/17. We carried out matching for this group using the NN and CEM matching method. The results are presented below.

Join year	Method	Lower CI	Estimate	Upper CI	Sample size
Recent	NN	0.01	0.04	0.07	2,269
Recent	CEM	-0.01	0.02	0.05	2,298
All years	NN	-0.03	-0.02	-0.01	37,241
All years	CEM	-0.04	-0.03	-0.02	27,005

The estimated impact for new joiners is higher than that for all users, particularly when using NN matching; it goes from a small significant negative result to a small significant positive. This may mean that there is less impact for some of the longer term users of the SAM Learning platform than there is for new joiners. It may be that newer joiners are more motivated to engage with the product effectively or are more aware of newer features than some long term users.

7.2.1.2 Omission of historical variables

Looking at recent joiners also allowed us to address another issue from the main analysis. Variables relating to the historical performance of a pupil's school before joining SAM Learning would usually be an important aspect of the matching process. However, they were omitted in the main analysis because SAM Learning had a high proportion of very long term participants, and because in some cases there were errors in the dataset relating to school join year. Looking at recent joiners means we entirely avoid the first of these two problems, and the second can be avoided by omitting the small number of recent joiners with ambiguous join dates. Therefore, we were able to include variables relating to historical school performance when matching this group of pupils. We carried out matching both with and without these variables, using the NN and CEM matching methods; comparing the results from both approaches gives an indication of how much the omission may have affected the original results. The results are presented below.

Matching variables	Method	Lower CI	Estimate	Upper CI	Sample size
With historic	NN	-0.01	0.02	0.05	2,269
With historic	CEM	0.09	0.16	0.24	499
Without historic	NN	0.01	0.04	0.07	2,269
Without historic	CEM	-0.01	0.02	0.05	2,298

The estimated impact obtained from matching with historical school performance variables is similar to the original estimates when using NN matching, but considerably higher when using CEM. However, in order to obtain acceptable balance using CEM, we were forced to considerably reduce the sample size. Because of this, we would deem the CEM result to be less reliable.

Nevertheless, given that there are some substantial differences in estimates, we would suggest that focusing analysis on relatively recent joiners would be a useful strategy for future evaluations. Omitting historical school performance may give misleading results; SAM Learning is less affected than some because its treated pupils and schools are fairly similar to the general population, but this is not always the case, and some evaluations would suffer even larger differences in conclusions if these variables were omitted. However, matching based on historical performance data from ten or fifteen years ago is clearly undesirable too. Using recent joiners avoids this issue, and also has the advantage of providing estimates based on the most recent implementation of the intervention.

7.2.2 Zero dosage pupils

As noted in section 5.2, it was common for pupils with a SAM Learning account to fail to use the platform at all. There is no way to determine whether there was an intention to treat these pupils; in some cases, schools set up accounts for all of the pupils in a cohort but only passed account details on to selected pupils for use. To understand more about this group, we carried out matching using NN and CEM for zero dosage pupils in each cohort from 2010/11 - 2017/18. The results are presented below.

Method	Year	Lower CI	Estimate	Upper CI	Sample size
NN	2011	-0.11	-0.10	-0.08	9,128
NN	2012	-0.13	-0.11	-0.10	9,662
NN	2013	-0.10	-0.08	-0.06	11,948
NN	2014	-0.09	-0.07	-0.06	12,182
NN	2015	-0.11	-0.09	-0.07	10,783
NN	2016	-0.10	-0.08	-0.07	11,809
NN	2017	-0.07	-0.05	-0.04	11,073
NN	2018	-0.11	-0.10	-0.09	17,018
CEM	2011	-0.14	-0.12	-0.11	6,189
CEM	2012	-0.12	-0.11	-0.09	6,079
CEM	2013	-0.11	-0.10	-0.08	8,097
CEM	2014	-0.12	-0.10	-0.09	7,966
CEM	2015	-0.14	-0.12	-0.10	6,615
CEM	2016	-0.10	-0.08	-0.06	7,820
CEM	2017	-0.09	-0.07	-0.06	6,931
CEM	2018	-0.10	-0.09	-0.08	11,615

For the zero dosage group, the estimated impact is both negative and significant for every cohort, and this is robust across both matching methods. This provides further evidence that there may be unobserved differences between the pupils that we are unable to account for by matching based on the

NPD. One possible unobserved difference could be the level of motivation; it may be the case that the zero dosage pupils are less motivated than pupils in other dosage groups, and this is why their outcomes are significantly lower than otherwise similar comparison pupils. If there was a difference in motivation, this suggests that there was an intention to treat at least some of the pupils in the zero dosage group. There may also be other unobserved differences between the groups.

It is also possible that the significant positive effects found for the high dosage group are at least partly due to unobserved differences. This is likely to be an issue for most evaluations looking at different dosage levels using a quasi-experimental design; any intervention that allows the user some choice on their level of engagement will, almost by definition, end up with more motivated pupils in the higher dosage groups. In the absence of a direct measure of motivation, this could be resolved through the use of an instrument that is predictive of engagement but unrelated to motivation. However, it is not always the case that such an instrument can be identified.