

PROJECT TITLE	Aspire to STEM
DEVELOPER (INSTITUTION)	STEM Learning
EVALUATOR (INSTITUTION)	RAND Europe
PRINCIPAL INVESTIGATOR(S)	Dr. Alex Sutherland
TRIAL (CHIEF) STATISTICIAN	Dr. Alex Sutherland
STUDY PLAN AUTHOR(S)	Dr. Yulia Shenderovich, Dr. Alex Sutherland, Dr. Kata Mihaly, Dr. Sashka Dimova

Study Planⁱ version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]	08/02/2019	[leave blank for the original version]

Design overview

Design type		Weighted difference-in-differences
Unit of analysis (school, pupils)		Schools – with all outcomes at school level
Number of Units (Intervention, Comparison)		206 intervention schools (92 primary and 113 secondary or middle deemed secondary schools, 1 all-through school); TBD control schools. Recruitment completed in September 2018.
Outcomes	primary	Pupil attainment in STEM: <ul style="list-style-type: none"> - Key Stage 2 (KS2) attainment in mathematics - KS4 attainment in mathematics and STEM subjects
	secondary	Pupil progression: <ul style="list-style-type: none"> - KS5 enrolment in mathematics and STEM Teacher attrition: <ul style="list-style-type: none"> - Teacher leaving the school (turnover) - Teacher leaving the profession (wastage)
Outcome sources (instruments, datasets)	primary	School performance data
	secondary	School Workforce Census

Contents

Study Plan version history	1
Design overview	1
List of abbreviations	3
Intervention	4
Study rationale and background.....	8
Impact evaluation	9
Research questions.....	9
Participants	9
Outcomes.....	10
Study design	17
Further analyses	25
Secondary outcome analyses	25
Subgroup analyses.....	25
Treatment effects in the presence of non-compliance.....	25
Effect size calculation.....	26
Implementation and process evaluation	26
Cost evaluation	32
Ethics	32
Data protection.....	32
Personnel	33
Risks	34
Timeline (indicative)	35
Acknowledgements	36
References.....	37

List of abbreviations

AtS	Aspire to STEM
CPD	continuing professional development
DfE	Department for Education
DiD	difference-in-differences
EAL	English as additional language
FSM	free school meals
GBM	generalised boosted regression models
IoP	Institute of Physics
IPW	inverse probability weighting
KS 2/4/5	Key Stage 2/4/5
KS statistic	Kolmogorov-Smirnov statistic
LAD	Local Authority Districts
OA	Opportunity Areas
SMD	standardised mean differences
STEM	science, technology, engineering and mathematics
SUTVA	Stable Unit Treatment Value Assumption
SWC	School Workforce Census
TIDieR	Template for Intervention Description and Replication
TWANG	Toolkit for Weighting and Analysis of Non-equivalent Groups

Intervention

Aspire to STEM (AtS) is a complex whole-school intervention aimed at improving primary and secondary school pupil outcomes in subjects related to science, technology, engineering and mathematics (STEM). AtS uses a model where clusters of between 3-10 schools in the same geographical area share resources and support to form a partnership, in order to build a sustainable 'community of practice' in STEM, adapted to their local needs and context.

As part of this intervention, each cluster of schools receives: (i) funding (£25,000) for bespoke continuing professional development (CPD) for teachers and leaders; (ii) career guidance for students; (iii) access to STEM inspiration and enrichment activities delivered by a network of STEM Ambassador volunteers; and (iv) access to online resources via the STEM Learning 'STEM Club infrastructure and eLibrary'. The intervention is designed to lead to improvements at three levels in the schools:

- (1) *Leaders* (improved leadership, new relationships with STEM employers, better engagement with parents, families and the community);
- (2) *Teachers* (increased confidence, motivation and competence in teaching STEM, better knowledge and ability to use real-life and industry contexts, improved teacher retention);
- (3) *Students* (raised aspirations and ultimately, improved STEM outcomes).

The following updated Template for Intervention Description and Replication (TIDieR) framework for the intervention has been agreed on as part of the initial IDEA workshop in April 2018, and revised further in ongoing communication with the implementation team.

1. BRIEF NAME

Aspire to STEM (AtS)

2. WHY: RATIONALE, THEORY AND/OR GOAL OF ESSENTIAL ELEMENTS OF THE INTERVENTION

Aspire to STEM tackles low pupil aspiration through improved STEM teaching. The intervention team believe that improved STEM teaching more than anything else can improve pupil opportunity and drive social mobility. Via improved STEM teaching, leadership and community engagement, the intervention aims to raise aspirations and career awareness among students, and ultimately improve student STEM academic outcomes. See Figure 1 for a Theory of Change.

3. WHO: RECIPIENTS OF THE INTERVENTION

The intervention primarily targets teachers and leaders of STEM subjects and school senior leaders.

Aspire to STEM is being implemented as a part of the Teaching and Leadership Innovation Fund (TLIF), a 3-year initiative in England to support high-quality professional development for teachers and school leaders.¹ The programme will see STEM Learning support 40 partnerships of schools located either in 12 areas designated as Opportunity Areas (OAs) or in Local Authority Districts (LADs) which are rated lowest in England (identified as LADs 5 or 6). OAs were announced in 2016-2017 as areas "identified as the most challenged when it comes to social mobility" by the Department for Education (DfE)². These areas were identified drawing on the Social Mobility Index and the Achieving Excellence Areas Index (DfE, 2017: 1). The Opportunity Areas include 12 locations: West Somerset, Norwich, Blackpool, North Yorkshire coast, Derby, Oldham, Bradford, Doncaster, Fenland and East Cambridgeshire, Hastings, Ipswich, and Stoke-on-Trent. LADs rated 5 and 6 by DfE (within 1 to 6 range) are

¹ <https://www.gov.uk/guidance/teaching-and-leadership-innovation-fund>

² <https://www.gov.uk/government/news/social-mobility-package-unveiled-by-education-secretary>

local authority districts with the lowest scores on composite indicators that include measures of pupil attainment, such as Progress 8 scores, and “capacity to improve indicators” identified by DfE, such as the number of teacher trainees per 10,000 pupils (DfE, 2016: 10-11).

To be eligible for TLIF, schools in these areas have to be rated ‘requires improvement’ or ‘inadequate’ (3 or 4) in their most recent Ofsted inspection.³ Schools rated as ‘good’ schools can join the programme on DfE approval of the rationale.⁴ A number of schools in the eligible areas did not have an Ofsted rating at the time of recruitment, for instance because the previous school was closed and became an academy, and these schools were considered eligible as well. For the purposes of this study, the schools were considered enrolled in the programme (intervention arm) based on the schools joined by 16 October 2018 (“quasi-randomisation” date; Anders, 2017).

4. WHAT: PHYSICAL OR INFORMATIONAL MATERIALS USED IN THE INTERVENTION

Participating schools will be offered a variety of face-to-face and online CPD activities, complemented by teaching resources identified by the schools with the support of educational leads. Additionally, the programme targets pupil aspirations through targeted inspiration and enrichment activities (see Figure 1 for examples). Both physical and informational materials are tailored to individual partnership needs.

5. WHAT: PROCEDURES, ACTIVITIES AND/OR PROCESSES USED IN THE INTERVENTION + 6. WHO: INTERVENTION PROVIDERS/IMPLEMENTERS

Delivery will be implemented by STEM Learning quality assured providers, including the National Stem Learning Centre subject expert team, STEM Learning’s Network of Science Learning Partnerships, external educational experts and trusted partner organisations (Maths Hubs, Institute of Physics), as well as ‘STEM ambassadors’ who are STEM professionals registered and trained by STEM Learning.

Each partnership will be supported by an educational lead, an education expert who will help schools with needs analysis and action planning and then help the partnership plan and implement the required activities.

7. HOW: MODE OF DELIVERY

Online and face-to-face

8. WHERE: LOCATION OF THE INTERVENTION

Various – on-site at school or a partnering school, Science Learning Partnership or regional centre and at the National STEM Learning Centre

9. WHEN AND HOW MUCH: DURATION AND DOSAGE OF THE INTERVENTION

Approximately 2-year support package, ending March 2020. Each partnership will have access to £25,000 of support. The amount is allocated by schools collectively, primarily to cover costs of CPD. There is no rule that schools will spend a specific proportion of the amount, e.g. equal amounts of the total. The money is overseen by STEM Learning who have to approve the expenses.

To ensure the feasibility of intervention roll-out, schools were recruited in two cohorts, and there are some differences between the timing of recruitment in the two cohorts.⁵ Namely, recruitment started approximately one school term earlier in Cohort 1 (Spring 2018 for Cohort

³ <https://www.gov.uk/guidance/teaching-and-leadership-innovation-fund-programmes-for-teachers-and-school-leaders#eligibility>

⁴ The evaluation will record the Ofsted ratings for schools included in the study.

⁵ Schools in the same partnership all belong to the same cohort

1 versus Autumn 2018 for Cohort 2) but implementation ends for *both cohorts* in March 2020 (see Figure 2). Thus, the programme lasts under two years for Cohort 2 (Sept 2018 – March 2020). In practice, however, some Cohort 2 schools moved through the implementation process faster and started CPD earlier than some Cohort 1 schools. Therefore, schools from both Cohorts are analysed jointly in this evaluation.

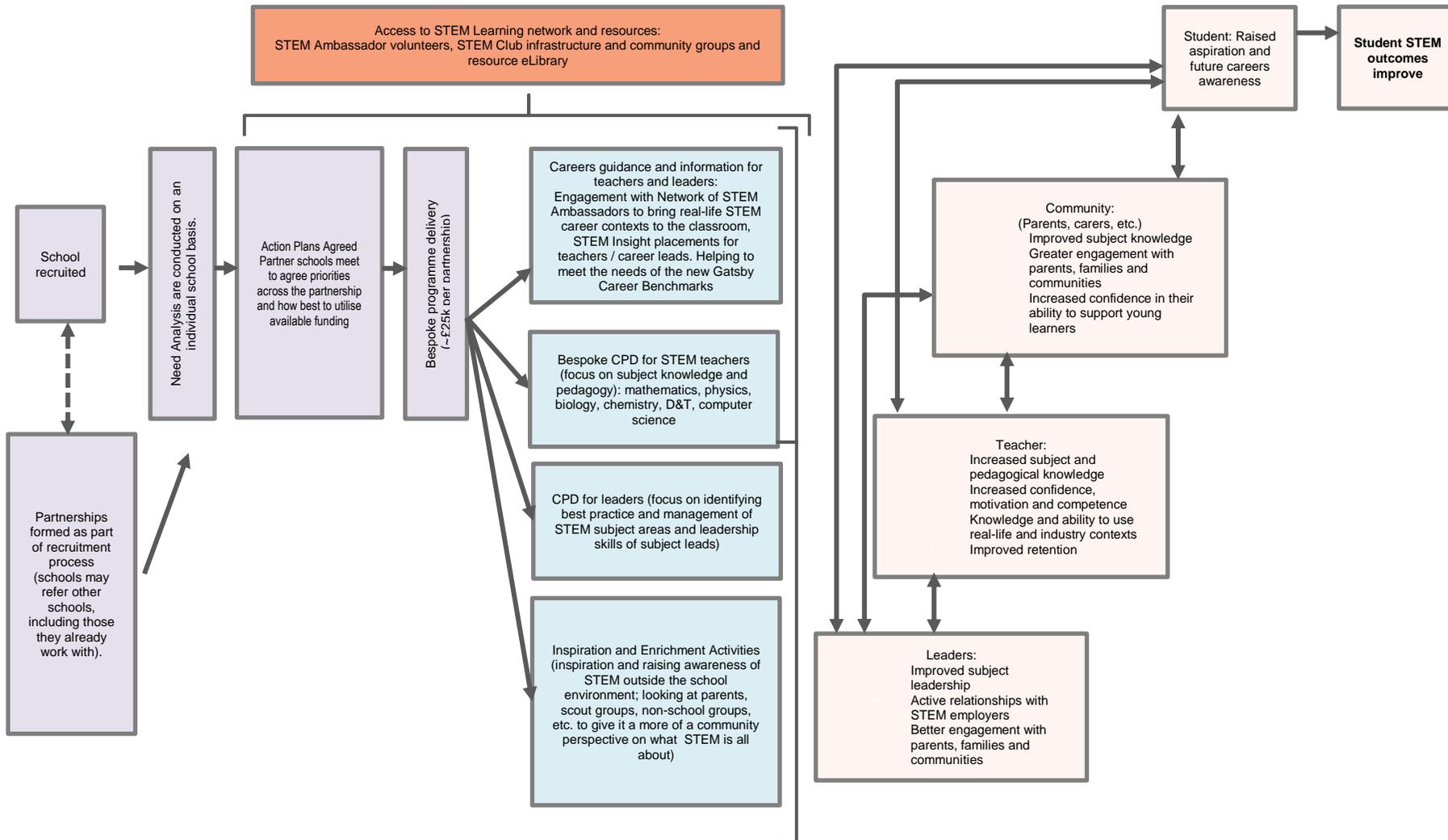
10. TAILORING: ADAPTATION OF THE INTERVENTION

Individual schools will have undergone a needs analysis with the help of the STEM Learning educational lead in order to identify their bespoke support needs. Following this, partnership leads agree on an action plan for the partnership. Resulting in a combination of individual school and a wider partnership support package. It is envisaged that each partnership (40 in total) will have a focus specific to its STEM-related needs.

11. HOW WELL (PLANNED): STRATEGIES TO MAXIMISE EFFECTIVE IMPLEMENTATION

STEM Learning have delivered school centred partnerships for a number of years with great success. Using this model, and acknowledging potential concerns due to the nature of the target audience, each partnership will be allocated a dedicated educational lead. Both the schools and educational leads will have access to a number of support and assistance services.

Figure 1. Intervention logic model



Study rationale and background

Continued professional development (CPD) is regarded as a “key lever for improving teaching” (Opfer, 2016: 3) in existing literature. Although robust evidence is still limited, a recent randomised trial of professional development in Australia found a statistically significant improvement in observer-rated teaching quality following a pedagogy-based, collaborative professional development intervention (Gore et al., 2017). It has been recommended that sustained, active and collaborative CPD is most likely to lead to changes in teacher practice and pupil outcomes (Opfer, 2016). However, the evidence in this area also has important methodological limitations (Sims and Fletcher-Wood, 2018), and several CPD programmes with the characteristics described above did not demonstrate changes in pupil outcomes (e.g. Garet 2011; 2016).

Aspire to STEM is being delivered by a well-established organisation for the provision of school-led STEM CPD – STEM Learning. STEM Learning have a wide reach in England – for example, between 2010/11 and 2012/13, at least one teacher from 83% of all secondary schools has attended a course organized by the National STEM Learning Network (Allen & Sims, 2017). In total, between 2010/11 and 2012/13, 25% of science teachers in England attended at least one STEM Learning course (Allen and Sims, 2017). Although some of the previous STEM Learning CPD initiatives have been evaluated, they have not been evaluated with rigorous counterfactual designs focusing on changes in pupil outcomes, such as the one proposed here.⁶ This research thus has the potential of filling an important evidence gap on the effectiveness of CPD and enrichment programmes in influencing pupil STEM learning.

Furthermore, Aspire to STEM may be able to contribute to addressing the challenges of training and retaining STEM teachers (Bélanger and Broeks, 2016). Research in the US suggests that high teacher turnover negatively affects student attainment (e.g., Atterberry et al, 2016), and teacher turnover and shortages are also a concern in the UK, especially as pupil numbers are increasing (Foster, 2018). While STEM teachers are more likely than non-science teachers to leave the profession, CPD may help increase their retention (Allen and Sims, 2017). A study by the Education Datalab (Allen and Sims, 2017), drawing on the School Workforce Census 2010-2015 (SWC), found that for secondary school STEM teachers, participation in STEM Learning CPD was associated with an increase in odds of staying in the profession by 160 percent.

Another set of estimates, at the department level, suggested that science departments where at least one teacher went on a STEM Learning training course had a decrease of 4 percentage points in the proportion of teachers leaving the profession in the two years following CPD (Allen and Sims, 2017: 9). The authors suggest that this is a substantial difference, as the science departments where no staff participated in the courses had around 10% of teachers leaving the profession per annum. It is important to note that while the teachers who attended CPD were matched with comparison teachers on a range of teacher demographic and career characteristics as well as school characteristics, there might be unobserved differences between teachers who attended CPD and those who did not that also influence the differences in leaving the profession. The relation of STEM Learning CPD to retaining teachers in their school (as compared to moving to another state school) was less conclusive in the study, with the most rigorous models not identifying any impact of CPD attendance on teacher retention within the same school.

Primary and secondary schools

⁶ See evaluation reports at <https://www.stem.org.uk/impact-and-evaluation>

Aspire to STEM includes both primary and secondary schools. There are differences between how the intervention operates in primary and secondary schools. In primary schools, it is common for teachers to be generalist and degrees in STEM subjects are not expected. In secondary schools, teachers are expected to have relevant education and/or experience to be able to teach STEM subjects. Therefore, teachers in primary and secondary schools differ in the type of STEM CPD support they need. Given the differences between science teaching and, as a result, Aspire to STEM delivery, in primary and secondary schools, we can think of the Aspire to STEM in primary and secondary schools as two versions of the programme. This suggests – as we discuss below – a need to conduct analyses separately for primary and secondary schools. Furthermore, a combined average effect would not be applicable to either primary or secondary schools.

Impact evaluation

Research questions

This study attempts to explore outcomes for schools that implement Aspire to STEM compared to schools running business as usual in terms of STEM CPD and enrichment. The research questions this study addresses are:

Impact questions:

1. Does Aspire to STEM lead to improvements in pupil attainment in KS2 Maths, KS4 Maths and other KS4 STEM subjects compared to 'business as usual' (in primary or secondary schools)?
2. Does Aspire to STEM lead to a higher proportion of students opting to take STEM subjects at KS5, compared to 'business as usual' (in primary or secondary schools)?
3. Does Aspire to STEM lead to fewer (STEM) teachers leaving their *school* (**teacher turnover**), compared to 'business as usual' (in primary or secondary schools)?
4. Does Aspire to STEM lead to fewer (STEM) teachers leaving *the profession* (**teacher wastage**), compared to 'business as usual' (in primary or secondary schools)?

Questions 3 and 4 refer to all teachers in primary schools and STEM teachers specifically in secondary schools – see “Teacher outcomes” section below for the definition of a STEM teacher.

Implementation and process evaluation questions:

5. How well was Aspire to STEM implemented in schools?
6. What are the typical components of Aspire to STEM?
7. What are the barriers and enablers for implementation of Aspire to STEM?

The evaluation of Aspire to STEM will be conducted at school level, using administrative data. The evaluation will focus on comparing change over time in outcomes in the schools that enrolled in Aspire to STEM to change over time in eligible schools that did not enrol in Aspire to STEM (see Study Design section below for more details).

All intervention schools enrolled on the quasi-randomisation date will be included in the analyses as intervention schools, even if they later drop out from the intervention. Schools may join the programme after this date, but only those schools already participating at the quasi-randomisation date will be included in the treatment condition for the evaluation. Similarly, for the main analyses all comparison schools will be included based on the status at the time of quasi-randomisation.

Participants

Aspire to STEM is a whole-school intervention. Inclusion/exclusion criteria for schools are defined above. To facilitate creation of partnerships, STEM Learning prioritised recruitment by

examining Google Maps to focus on areas where many eligible schools are located in proximity of each other, meaning that isolated schools otherwise eligible would not be approached because of location. This has the benefit of offering a means to create alternative comparison groups (discussed below). The recruitment process also aimed to keep all schools within an education authority together as these schools are likely to have existing collaborations.

Outcomes

Timelines

The outcomes for the study reflect both the aims of the intervention and the anticipated long-term benefits for pupils arising from having more skilled/better resourced STEM teachers and leaders working in schools. The timelines for anticipated outcomes were selected in consultation with STEM Learning based on their experience. One challenge for the current evaluation is that whole-school changes may take a longer time to become visible. As currently designed, the research team will receive the results on KS2 and KS4 outcomes in December 2020 – January 2021, based on pupil performance in the spring/summer of 2020 (see Table 1 and Figure 2 below for details on the timeline). The teacher retention and KS5 outcomes would be reported later as an addendum to the original report.

Cohorts 1 and 2

Outcomes from the same academic years are considered for all schools to maximise comparability. As there was a lot of variation in school implementation timelines, for instance some schools in Cohort 1 finalising their action plans later than schools in Cohort 2, we have decided to analyse both cohorts jointly.

Table 1. Timeline of anticipated school-level outcomes

AtS impact timeline	Measure	Collected by DfE	Outcome measure available for analysis (approximate)
Primary outcome: Student attainment in KS2 Mathematics and KS4 Mathematics and STEM subjects based on 2019/2020 (data available online)	KS2 2019	May 2020	December 2020
	KS4 (GCSE)	June 2020	January 2021
Secondary outcome: STEM take up - number of pupils entered in STEM A-levels in 2018/2019 (data available online after A-levels are taken)	KS5 (A-level) ⁷ 2021	June 2021	March 2022
Secondary outcomes: - Teacher wastage (leaving the profession) - Teacher turnover (leaving the school for another government school) SWC data to be requested from DfE)	School Workforce Census 2020	November 2020	November 2021 ⁸

Pupil outcomes

As clarified above, to facilitate reliance on administrative data all outcomes will be considered at the school level. It was agreed to focus only on measures based on standardized Key Stage tests rather than teacher assessments. Therefore, as standardized assessment is not available for science outcomes in KS2, only mathematics is included as part of KS2 assessment. For pupil attainment in secondary school, we will draw on KS4 (GCSE) results. As there are no ready-made variables reflecting all relevant STEM subjects, we propose to combine the scores from all relevant subjects (see Table 2), obtaining a STEM-subject GCSE school average (mean).⁹

Due to changes in GCSE (KS4) grading from 2017/18, KS4 outcome would be standardised (z-scored) to ensure comparability across years for the analyses. We have also considered including alternative outcome measures, such as Progress 8 for Key Stage 4. However, these measures compare progress to average change in all English schools, while our primary study focus is comparing programme effects to change in schools in the Opportunity Areas.

⁷ If possible, AS-level progression will be considered as well.

⁸ The date depends on when DfE can share data as the teacher-level data are not available online (and school-level data on retention are not prepared by DfE)

⁹ Note that while KS5 outcomes and teacher turnover/wastage are included in the outcome list, given the timeline they would be included in a later addendum rather than the primary evaluation report.

At KS5, recent reforms mean that AS grades do not contribute to A-level grades, but it is not clear yet how this will affect the number and type of pupils who enrol in AS-levels and sit AS-level exams. The evaluation of impact on progression to KS5 STEM subjects will focus on enrolment in A-levels, but will also aim to incorporate AS-levels progression for the same cohort of pupils. However, we will not combine AS- and A-level data.

While drawing on administrative data offers a practical approach to including many schools in the evaluation, one limitation of the approach is that, focusing on the school level results, we are not able to account for student turnover and examine to what extent the results are influenced by composition effects. Ideally, the analyses would control for prior attainment of the same cohort of students rather than controlling for earlier attainment at that age in the same school.

Teacher outcomes

We focus on whether the intervention affects the percent of teachers staying in the same school. As mentioned above, teacher may leave either due to going to another school (turnover) or due to leaving the profession altogether (wastage). As the Census does not include private schools, moving to them is considered equivalent to leaving the teaching profession as the teacher is no longer present in the Census.

All schools start implementation between two rounds of the Census (November 2017-November 2018). We propose to define teacher wastage so that the wastage variable for school m in period t is defined as the proportion of teachers who were teaching in period $t-1$ in that school and are no longer teaching in any state school in period t . Similarly, teacher turnover is the proportion of teachers who are no longer teaching in the school m in period t . Specifically, our secondary outcomes analysis will examine whether or not teachers have left the school in Autumn 2019 (based on Census in November 2019), compared to 2018-2019 academic year (Census in November 2018). This definition allows for consistency with baseline weighting variables (wastage/turnover in previous year). We will aggregate the teacher-level information on wastage and turnover to the school level.

In addition, it is important to consider the definition of a “STEM teacher”. We propose to follow the definition used by Allen and Sims (2017), where in secondary school, a science teacher is someone who in the focal period (in this case, during Aspire to STEM recruitment period, academic year 2017/2018) has taught science for greater than or equal to half of their timetabled teaching hours, and spent at least one hour a week teaching science. Thus, if a teacher taught for 2 hours a week, one of which was science they would be considered a STEM teacher in this study.

Several caveats apply to using the SWC. There is a considerable amount of missing data in the Census at subject level, so for subjects not included into the top-level subject of “science”. For subjects such as computer science, relevant following a definition of STEM in this evaluation, there is likely to be missing data leading to smaller number of STEM teacher identified (Sam Sims, personal communication, 17 September 2018). Furthermore, teachers who have not been yet put on the official payroll at the time when Census data are submitted, would not be included in the Census (for instance early career teachers). As data are provided at teacher level, our analyses will require generating school-level values for wastage and turnover.

In primary school, as all teachers are involved in teaching STEM subjects, all teachers would be included in respect to retention. As mentioned already, given the noteworthy differences between how the intervention operates and how the outcomes are defined in primary and secondary schools, we propose to explore effects separately for both types of schools (see “Outcomes” below).

Figure 2. Timeline for implementation and study outcomes

		Matching data 2015/16, 16/17	2017/18			2018/19			2019/20			2020/21			2021/22			2022/23	
			Autumn	Spring	Summer	Autumn	Spring	Summer	Autumn	Spring	Summer	Autumn	Spring	Summer	Autumn	Spring	Summer	Autumn	Spring
		COHORT 1		Recruitment															
					Implementation														
		COHORT 2			Recruitment														
					Implementation														
Data collected:	Data available for outcome analysis:																		
KS2 & KS4 (May-June)	✓ (December/January)								✓ KS2/4 data from 2018/19 (test year)	KS2 & KS4			✓ KS2/4 data from 2019/20 (outcome year)						
KS5 A Level enrollment (June)								KS2 & KS4 2018/2019 Enrollment in KS5 for 2021											✓ Data for those enrolled in KS5 in
KS5 exam	✓ (March)												KS5 exams taken						
Teacher retention (November)	✓ (November)											Retention							✓ Retention 2020/21

Note: In **bold** are the time points used for outcome analyses

Table 2. Primary/Secondary Outcomes and Data Sources

Outcome type	Source dataset & years	Outcome	Specific variable(s)
Primary	KS2 Mathematics (school level datasets released by DfE) ¹⁰	Pupil attainment in primary school (school average)	PTMAT_EX - % of pupils achieving the expected standard in maths at KS2 (test-based) ¹¹ . See Appendix 1 for additional details on KS2 Maths variables.
Primary	KS4 Mathematics and KS4 Science GSCE subjects ¹²	Pupil attainment in secondary school (school average)	Average GSCE grade across the STEM subjects ¹³ OR Percentage of pupils with grades 9-4 in STEM subjects.
Secondary	KS5 take-up of STEM A-level subjects	Pupil progression (school average)	“Total entries” for all STEM subjects (divided by) TPUP1618 - “Number of 16-18 year old students in sixth form”
Secondary	School Workforce Census 2019	STEM teacher attrition (school average): 1. Wastage 2. Turnover	- SchoolYear - StaffMatchingReference - QualifiedLeaverType - Subject code - Hours

¹⁰ <https://www.compare-school-performance.service.gov.uk/download-data>

¹¹ Note that an alternative approach would be to use MAT_AVERAGE, scaled score in maths – see Appendix 1 for details.

¹² Science subjects include all relevant science subjects, such as Mathematics, Biology, Chemistry, Computer Science, Computer Technology, D & T Engineering, D & T Product Design, D & T Resistant Materials, D & T Textiles Technology, Electronic Engineering (General), Electronics (Physics), Engineering Studies, Physics (General), Science (General/Combined), Statistics

¹³ Data are arranged by subject, with the number of pupils who earned each grade available.

Sample size calculations

Based on STEM Learning school recruitment lists, we present the numbers of eligible and enrolled schools in the tables below (Tables 3 and 4).

Table 3. Overview of eligible schools

Type of school	Number in STEM Learning list
Primary schools	1,007
Secondary schools (incl. middle schools deemed secondary)	392
Secondary special schools ¹⁴	1
All-through schools (including both primary and secondary)	9
Total	1,409

Table 4. Overview of Aspire to STEM enrolled schools

Type of school	Number in STEM Learning list
Primary schools	92
Secondary schools (incl. middle schools deemed secondary)	112
Secondary special schools	1
All-through schools (incl. both primary and secondary)	1
Total	206

We calculate MDES separately for primary and secondary schools as intervention effects will be presented separately. This is because a combined (averaged) effect would not be applicable to any school if one mixes primary and secondary schools together.

Primary schools

There are few clear guidelines on power calculations for quasi-experimental designs (Hedges, 2017). We base the minimum detectable effect size calculation for primary schools on two scenarios:

1. A 1:1 allocated individually randomised controlled trial, 184 schools in total (92 schools in treatment and control respectively).
2. A 1:9 allocated individually randomised controlled trial, treatment group of 92 schools, comparison group of 828 (10% of schools allocated to treatment, with 9 comparison schools compared to one treatment school), 920 schools in total.

These scenarios are based on the numbers of eligible schools and enrolled intervention schools, provided by STEM Learning, as presented in Tables 3 and 4. Namely, from 1,007 eligible primary schools, 92 were enrolled, leaving up to 915 schools for comparison. Increasing the number of schools increases precision of estimates, and we have used the minimum and maximum to illustrate the impact that different scenarios might have.

¹⁴ Special schools will not be included into the study.

Table 5 compares scenarios 1 and 2. In the first scenario, using power calculations for an individually randomised trial with a 1:1 allocation ratio as a benchmark, the minimum detectable effect size would be $d = 0.322$. This represents the ‘worst case scenario’ as we anticipate being able to include more schools in the analysis. In the second scenario, a 1:9 allocation ratio, the minimum detectable effect size would be $d = 0.239$. This is an optimistic estimate of MDES because not all schools would contribute equally in the weighted design even if they could all be included, but it provides a justifiable ‘best case scenario’. However, given the weighting approach, the effective sample size and MDES may be lower. Assuming a 25% reduction in the effective sample size, MDES for primary schools may lie between 0.276 and 0.372.

In both scenarios we use the following parameters: an alpha of 5%, desired power of 80%, 40% of variance in the outcome explained by covariates. Power calculations were completed using PowerUp! (Dong and Maynard, 2013).

Table 5. Minimum detectable effect sizes based on proposed study design for primary schools

	Study Plan	
	Allocation ratio	
	1:1 (worst case)	1:9 (best case)
MDES	0.322	0.239
Percent of variance in outcome explained by covariates	.40	.40
Alpha	0.05	0.05
Power	0.8	0.8
One-sided or two-sided?	Two	Two
Number of schools	intervention	92
	control	828
	total	184
		920

MDES is based on two individually randomised controlled trials, one with 1:1 allocation and the other with 1:9 allocation, based on the number of eligible primary schools.¹⁵

While the number of pupils does not directly affect our ability to detect intervention effects in this study, based on average school numbers in 2015/16 (Jackson, 2017), we note that approximately 23,920 children could be affected by Aspire to STEM intervention (as 260 pupils attend an average primary school, with 92 intervention schools in the programme). Adding in the potential comparator schools means the number of pupils whose schools are included in the evaluation would be at least doubled – and in practice would be many times more (e.g. 1:9 matching would ~250,000 pupils).

Secondary schools

Similarly, Table 6 compares scenarios 1 and 2 for secondary schools, based on the numbers of eligible and recruited schools. In the first scenario, using power calculations for an

¹⁵ We do not propose to correct for multiple outcome testing as we are conducting outcome analyses on two different groups (populations: primary and secondary schools).

individually randomised trial with a 1:1 allocation ratio as a benchmark, the minimum detectable effect size would be $d = 0.291$. In the second scenario (1:2 allocation), the minimum detectable effect size would be $d = 0.253$. As for primary schools, these scenarios are based on the numbers of eligible schools presented in Tables 3 and 4. Out of 392 secondary schools (including middle schools deemed secondary), 112 were enrolled in Aspire to STEM, leaving over 200 schools for comparison. Assuming a 25% reduction in the effective sample size due to the weighting, MDES for secondary schools may lie between 0.292 and 0.337.

Table 6. Minimum detectable effect sizes based on proposed study design for secondary schools

		Study Plan	
		Allocation ratio	
		1:1	1:2
MDES		0.291	0.253
Percent of variance in outcome explained by covariates	school-level	.40	.40
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		Two	Two
Number of schools	intervention	112	112
	control	112	224
	total	224	336

MDES is based on two individually randomised controlled trials, one with 1:1 allocation and the other with 1:2 allocation, based on the number of eligible secondary schools.

With an average secondary school size of 910 pupils (Jackson, 2017), the 112 interventions schools would include roughly 101,920 pupils.

Study design

We propose a difference-in-differences design, using inverse probability weighting (IPW) on the propensity score to improve the suitability of comparator schools. The steps in our analysis are: (1) choosing the primary treatment effect of interest; (2) estimating the propensity score; (3) evaluate the quality of weights; (4) estimate treatment effects.

In the sections below, we clarify our rationale for the selection of these methodologies. Briefly, propensity score is an individual's (or unit's) probability of receiving treatment or exposure given the observed pre-treatment characteristics (Williamson et al., 2011). A key motivation for this approach is that many variables are reduced to a single score. Weighting on the inverse of the propensity (or probability) score assigns a weight to each observation, based on the estimated probability of this unit's exposure to treatment. Weighting aims to produce a (potential) comparison group that has the same distribution of observed baseline characteristics as the intervention group (Williamson et al., 2011). The difference-in-differences approach compares the difference in outcomes between the study groups before and after the intervention. Overall, the combination of difference-in-difference and IPW allows investigating programme effects over time, while generating comparable comparison groups (Stuart et al., 2014).

Table 7 provides an overview of main study design features, with further details described in the sections below.

Table 7. Overview of study design

Item	Response
Level of comparison and analysis	School
Evaluation approach	Difference-in-difference combined with weighting on the inverse of the propensity score.
Weighting variables	See Tables 8 and 9
Sensitivity analyses	See section “Robustness checks and identification assumptions”

We will separate the outcome variables until we have assessed the quality of the weights, generated as described above. We will not incorporate the outcome data into the model until the weighting model has been developed. This separation is facilitated by the fact that outcome data will not become available until 2020.

Propensity score calculation

We will use publicly available data on school characteristics (see Tables 8 and 9 for details) to produce propensity scores to generate a comparison group for treatment schools based on observable data (Rosenbaum and Rubin, 1983).

The propensity score is defined as the ‘conditional probability of being exposed [to an intervention] given the observed covariates, $e(X) = P(Z = 1 | X)$ ’ (Williamson et al., 2011: 278). Propensity scores can be calculated in different ways, but a well-known approach is through a logistic regression:

$$\text{Logit}(\pi) = \alpha + \beta_1 X_1 + \beta_2 X_2 \dots \beta_k X_k + \varepsilon$$

Where $\text{Logit}(\pi)$ is the log-odds of the probability that the outcome (treatment Yes/No) is equal to 1, expressed as a linear function of the predictors $\beta_1, \beta_2 \dots \beta_k$ and (prediction error). The resulting prediction from this model, ranging between 0-1, constitutes the propensity score.

Calculation of the propensity score uses data to estimate the relationship between observed predictor variables, intervention, and outcomes (Guo and Fraser, 2010). School-level prior attainment, being in an Opportunity Area or LADs 5 or 6, and low STEM prior attainment are key variables to predict inclusion in AtS. We have carefully considered what variables to include based on prior research about STEM attainment (see Table 8). One must acknowledge that a limitation of propensity scores is that the method accounts only for observable data in the matching process, which is part of the motivation for combining PSM with difference-in-differences, as DiD accounts for all time-invariant unobserved factors (discussed below).

In addition, as much as possible we will include information on other relevant and recent interventions in the intervention and comparison schools, for instance previous intensive STEM Learning programmes and current CPD by the Institute of Physics delivered as part of TLIF. Institute of Physics (IoP) is the only other provider funded by TLIF Round 1 that targets STEM subjects (physics CPD), and both STEM Learning and IoP team coordinated their recruitment – however, a few schools are enrolled in both IoP and STEM Learning programmes. Furthermore, a later TLIF Round 2 involves other STEM-related providers.

The main analysis will be to compare AtS schools to schools that did not participate in AtS, with separate analyses conducted for primary and secondary schools. The ‘did not participate’ group will consist of different reasons for non-participation, which we discuss next, but the headline result from the study will be a comparison of AtS schools to all non-AtS schools, based on the quasi-randomisation date.

While we are able to account for a number of school characteristics, important unobserved variables such as the supportiveness of the senior leadership team may influence both a school’s decision to join Aspire to STEM and levels of attainment in a school, acting as an unobserved confounder. However, as suggested by Anders et al. (2017: 25), pre-intervention pupil attainment trends can also act “as a proxy for school motivation”.

Furthermore, there are several comparisons that could be made to account for potential differences in school motivation to engage in a demanding CPD programme that we will explore in sensitivity analyses. The proposed comparisons for this study are to compare AtS schools to:

1. Eligible schools not approached by STEM Learning, e.g. due to location, by the quasi-randomisation date (comparison 1)
2. Eligible schools approached by STEM Learning and refused/unable to participate, by the quasi-randomisation date (comparison 2)

As noted above, comparison groups 1 and 2 will be combined for the main analyses but explored separately in sensitivity analyses. Each comparison group has its potential shortcomings. Schools in group 1, not approached for recruitment, are likely to be more remote and geographically isolated than the intervention schools. Some of the schools in group 2 that have been approached by STEM Learning have instead taken part in other TLIF interventions, including the intervention offered by IoP. Thus, there will be some equally “motivated” schools in comparison 2 group, which did not participate because they were already taking part in the IoP Physics programme, or because they decided to prioritise literacy, for example.

Given the focus on the intervention in these geographic locations, choosing schools that could realistically have been included in the treatment are closer to ‘focal local controls’ (Campbell, 1976). In principle, looking at schools outside of the eligible areas – for instance, similar schools in proximity to OAs, which did not have a chance to enrol in *Aspire to STEM* – could also offer valuable information. However, it is not feasible within the scope of the current study. A key limitation on the number of comparator schools is the feasibility of data preparation as schools’ names and ID numbers change over time – while this information is available online, it requires manual checks that can be time consuming.

Weighting variable selection

The primary goal in selecting predictors is identifying variables that we think are predictive of either treatment or outcome, or both. However, we also propose to include predictors to help ensure comparability between intervention and comparator schools, for instance to account for different school types.

Existing reviews of literature suggest that disadvantage and literacy are important predictors of science attainment (Nunes et al., 2017). As the evaluation focuses on school-level outcomes, we therefore suggest drawing on school-level percentage of FSM pupils as a weighting variable. A recent systematic review of factors linked to poor academic performance of disadvantaged school students in science and maths (Banerjee, 2016) outlines other relevant factors, such as parental participation – however, these are not recorded at school level. There is some indication that single-sex and co-educational schooling could affect pupil attainment (Pahlke, Hyde, & Allison, 2014), so we include a predictor for same-sex (all girl or all boy) or a mixed-sex school.

Given the longitudinal nature of the intervention and the changes in school characteristics over time, it is important to define the relevant years for each predictor – we recommend using multiple years for pupil attainment and teacher retention. For the other predictors, we will only include information from the first year of programme recruitment, 2017/2018 academic year. Other than attainment measures, the variables used for weighting primary and secondary schools will be the same or equivalent.

Table 8. Proposed weighting variables for pupil outcomes

Variable(s)	Notes
School-level pupil academic attainment from past 2 years (2015/16, 16/17): <ul style="list-style-type: none"> - KS2 maths - KS2 reading - KS4 maths¹⁶ - KS4 science 	Constructed in the same manner as intervention outcomes
% FSM pupils (2017/2018)	Predictor of STEM attainment
% of EAL pupils (2017/2018)	
School type (maintained, academy, etc.) (2017/2018)	
Same-sex or mixed school (2017/2018)	
School size (2017/2018)	
Other interventions implemented by STEM Learning in the past 2 years (2015/16, 16/17)	STEM learning to provide a list of schools taking part in interventions of sufficient intensity to be comparable to Aspire to STEM – given the wide reach of this organization, it is important to account for their ongoing interventions in potential comparison schools.
Physics CPD by IoP as part of TLIF	

Table 9. Proposed weighting variables for teacher outcomes

Variable(s)	Notes
School-level pupil academic attainment from past 2 years (2015/16, 16/17): <ul style="list-style-type: none"> - KS2 maths - KS2 reading comprehension, spelling, punctuation and grammar - KS4 maths - KS4 science 	Constructed in the same manner as intervention outcomes
% FSM pupils (2017/2018)	Predictor of STEM attainment
% of EAL pupils (2017/2018)	
School type (maintained, academy, etc.) (2017/2018)	
Same-sex or mixed school (2017/2018)	
School size (2017/2018)	
Other interventions implemented by STEM Learning in the past 2 years (2015/16, 16/17)	STEM learning to provide a list of schools taking part in interventions of sufficient intensity to be comparable to Aspire to STEM – given the wide reach of this organization, it is important to account for their ongoing interventions in potential comparison schools.
Physics CPD by IoP as part of TLIF	

¹⁶ Due to the changes in the GCSE scoring from A*-U to 9-U, in 2016/17 Mathematics scores are available in the new grading system while other STEM subjects in the previous grading system, so the scores will be standardised before averaging.

Teacher attrition/retention level in the previous year (2016-2017)¹⁷

Next, we describe our approach for using the propensity scores to weigh observations.

Inverse probability weighting on the propensity score

Our preferred approach to generate a comparison group is IPW on the propensity score.¹⁸ Weighting is advantageous because it uses all available data rather than only matched schools; discarding data through PSM is one key criticism from King and Nielsen (2016) regarding matching on the propensity score (see also Dahabreh et al., 2012). Weighting also does not assume that all observations are equally important (as naïve difference-in-differences does). Schools “closest” or most similar to intervention schools in terms of their observed characteristics (pscore) are weighted more than schools that have less in common. Weighting also requires fewer decisions about the exact analysis approach, so can be pre-specified more readily, and there is a growing literature on ‘best practice’ for using IPW (e.g. Cefalu et al., 2015; Lockwood et al., 2016). However, while weighting allows to draw on more data, which should increase statistical power, it can still lead to high variance and, therefore, wide confidence intervals when units with very high or low propensity scores are present (Williamson et al., 2011).

To pursue IPW, we will use the RAND-developed Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG) to conduct the analysis (Griffin et al., 2014). There are three benefits to our using of TWANG. First, the TWANG algorithm, not the user, determines the most appropriate model, thereby reducing what Loken and Gelman (2017) call ‘researcher degrees of freedom’. Second, the programme has a number of user-friendly features to assess the quality of the propensity score weights, including statistical and graphical output. Lastly, the RAND research team can call on in-house expertise in case there are any estimation concerns. TWANG will produce a set of weights for the comparator schools that can be used in the analyses.

TWANG estimates pscores and weights using an automated, non-parametric machine-learning technique that requires the analyst to specify pre-treatment covariates to balance between the treatment groups (known as generalised boosted regression models, or GBM). Boosted regression models, as implemented in TWANG, are well-suited to optimise matching and weighting models, leading to better bias reduction and more stability (consistency) (Lee, Lessler, Stuart, 2010). Machine learning approaches, such as GBM, do not rely on a parametric logistic regression and therefore can “minimize bias from incorrect assumptions about the form of the model used” (Parast 2017: 14). Weights produced in similar approaches have been demonstrated to generate more precise estimates of treatment effect and lower mean squared error, compared to logistic regression (Harder et al. 2010; Lee et al, 2010).

This approach uses all the covariates included in the model, upweighting/downweighting those that are more/less important for prediction. The models also include higher order terms (interactions) and allow the use of multiple forms of predictors, e.g. continuous, binary, and automatically balances on missingness. Should there be any issues with the weighting using the TWANG estimation, we will draw on another estimator, such as the Covariate Balancing Propensity Score (CBPS; Imai & Ratkovic, 2014)

Difference-in-differences

Once the weights for schools have been identified, we will then undertake a weighted difference-in-differences analysis (see Wooldridge, 2002). The difference-in-differences (DiD)

¹⁷ In principle, this variable could also be used for weighting for pupil attainment outcomes but likely we will not have this data in time for the main report (this is planned for the addendum).

¹⁸ For average treatment effects on the treated (ATT) estimates, the treatment group is not weighted, but the control group is weighted using $\text{pscore}/(1-\text{pscore})$. For average treatment effect (ATE) estimation the weight is: $1/\text{pscore}$ for the treatment group and $1/(1-\text{pscore})$ for the control group (Williamson et al., 2012).

estimator compares the difference in outcomes before and after the intervention for intervention schools to the same difference for comparison schools, with each observation weighted according to the propensity score. This reduces bias in the post-intervention differences between the treatment and comparison schools that could be the result from time-invariant differences (both observed and unobserved) between these two groups. Time-varying factors, such as nation-wide changes in grading approaches, will affect both control and intervention schools similarly, so any difference in trends should be attributable to the intervention (i.e. the only time-varying variable that differentially affects one of the groups should be the intervention). Furthermore, multi-year data likely offers increases in statistical power by providing multiple observations of the school performance over time (McKenzie, 2012).

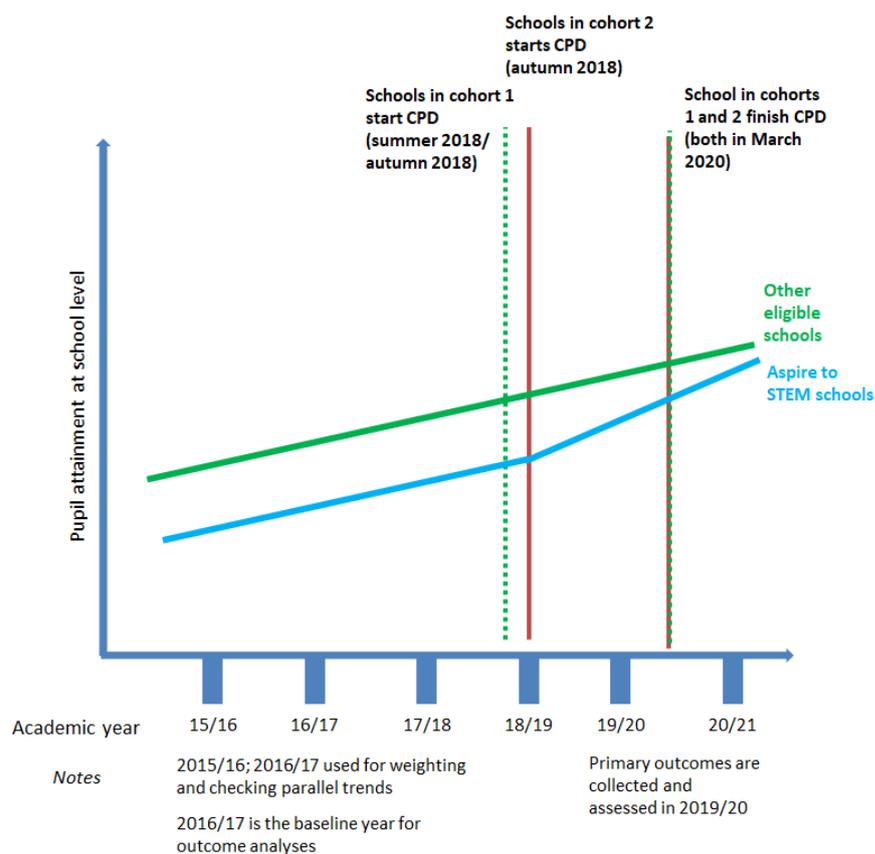
As any design, DiD relies on a number of assumptions, such as Stable Unit Treatment Value Assumption (SUTVA), meaning that the composition of intervention and comparison schools is stable over time and there is no spill-over (Stuart et al., 2014; Wing, Simon, Bello-Gomez, 2018). A key assumption of DiD, which we will examine visually for the two pre-intervention years, is that the difference between treatment and comparison schools is constant over time (i.e., parallel trends). For DiD it is not essential that pre-intervention trends are the same, merely that the trend lines are parallel. We will plot weighted observations and if they do not produce parallel pre-intervention trends, we will revisit the weighting model. As a comparison, we will also plot the non-weighted observations. For simplicity, we set out the DiD as if there were only two time-points. A DiD regression model can be represented as follows:

$$y_{st} = \beta_0 + TREAT_s\beta_1 + POST_t\beta_2 + TREAT_s * POST_t\beta_3 + X_{st}\beta_4 + u_{0s} + \varepsilon_{ts}$$

where y_{st} is the outcome for school s at year t ; $TREAT_s$ is a dummy that equals 1 for intervention schools and 0 for comparison schools in any time-period; $POST_t$ is a dummy that equals 1 for post-intervention and 0 otherwise; X_{st} is a vector of school-level characteristics that will be added to the model; and ε_{st} is an idiosyncratic error term. The coefficient of interest is β_3 , the interaction between intervention group membership and post-intervention outcome. This coefficient tells us about the difference in post-intervention outcomes, comparing intervention and control schools. School-level characteristics included as covariates will be based on the list of weighting variables to provide the “double-robust” estimation.

Figure 4 illustrates the comparisons that our approach facilitates. Note that the pre-intervention trends for all the schools are presented as if they were the same (here, for simplicity, constant). With analyses conducted separately for primary and secondary schools, the headline results will be (1) the overall effect size for primary schools and (2) the overall effect size for secondary schools.

Figure 3: Visualisation of proposed school-level comparisons



Assessing propensity scores and comparator quality

In addition to examining the parallel trends assumption, we will use the built-in approaches in RAND-developed Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG)¹⁹ for assessing baseline balance (Griffin et al., 2014; Cefalu et al., 2015 - TWANG is discussed in more detail below). Following the estimation of the propensity scores, TWANG will generate statistics assessing baseline balance in both table and graphic form. These will show “how well the resulting weights succeed in manipulating the control group so that its weighted pre-treatment characteristics match, or balance, those of the unweighted treatment group” (Ridgeway, 2017: 5-6).

We will examine all four available indicators of imbalance: namely, mean and maximum standardised mean differences between weighted pre-treatment variables as well as mean and maximum Kolmogorov-Smirnov (KS) statistic. The standardised mean differences for each variable are expressed as an effect size, standardised by dividing the difference means on a variable between the treatment and control groups by the standard deviation of that variable. While standardised mean differences demonstrate the average differences between groups, the KS statistic calculates the maximum vertical distance between groups across the cumulative distribution of a variable, where p-values are used to assess whether the groups are statistically different from one another. The KS adds value as it compares the entire distributions rather than only the means. However, unlike the SMD, the distribution of KS statistic is influenced by the sample size (McCaffrey et al., 2013).

To document imbalance after weighting, we will highlight any pre-treatment variables for which, after weighting, the KS *p*-value is less than 0.1 and the standardised mean difference

¹⁹ <http://www.rand.org/statistics/twang>

is greater than 0.2 (the thresholds suggested in McCaffrey et al, 2013). We will report on all the variables with remaining imbalance, if any. All pre-treatment variables (regardless of baseline equivalence) will be included in the regressions as part of the double robust approach (approach robust to misspecification of the propensity score model or the regression model, see Funk et al., 2011; Lunceford & Davidian, 2004).

Multi-year observations

We will plot the pupil attainment and teacher attrition data from each study year to observe trends over time and examine the pre-intervention trends (see page 23 on parallel trends assumption). For the outcome analyses, we propose reducing the number of time points as discussed below. This is based on considerations of the issue of serial correlation (Angrist and Pischke, 2009), the complexity of the study design (discussed in Study Design section on page 18), and the gradual introduction of the intervention across the intervention schools.

Specifically, in the outcome analyses, we plan to use the data from 2016/17 as baseline-, while 2019/20 will be used as the post-intervention period for the main pupil attainment outcomes (and 2020/21 for teacher outcomes). Data from 2017/18 and 2018/19 will not directly contribute to the outcome analyses as this is expected to be too early for the intervention to have an effect, but these years will be used for plotting the trends. As only the data from 2016/17 (average), and 2019/20 (or 2020/21, depending on outcome) will be used for the main analyses, we only consider the two time periods in the outcome models. The dependence between multiple observations per school in the analyses will be handled by multi-level modelling (observations clustered in schools).

Robustness checks and identification assumptions

As per Anders et al. (2017) we would assess sensitivity to model specification by re-running analyses using different specifications.

1. Comparing the intervention group to alternative control groups (see above in “Study design” for details).
2. Running a single-level outcome regression with pre-intervention values as covariates.
3. Using a modified set of variables in weighting schools (such as Attainment 8 or Progress 8 and FSM pupil attainment instead of Maths and English).
4. Placebo test in the year prior to treatment.
5. A final sensitivity analysis would be to assess susceptibility of results to hidden selection bias using Rosenbaum bounds (see Rosenbaum, 2002; Becker and Caliendo, 2007; Kirk and Sampson, 2013).

Further analyses

Secondary outcome analyses

Secondary outcome analyses will focus on teacher retention and take-up of STEM subjects at Key Stage 5, as specified above. Overall, the analyses will mirror those for primary outcomes.

Some difference will be introduced in the list of weighting variables for teacher outcome analyses, as described in Table 9. However, our ability in this study to include many of the predictors known to be relevant to teacher attrition is limited as there is no straightforward way to include person-level predictors, such as teacher's degree, working conditions, etc. given the analysis at the school level.

As the school is the unit of analysis, all outcome measures are continuous (average scores or percentages per school).

Subgroup analyses

See the clarification above regarding separate analyses in primary and secondary schools under "Outcomes". Furthermore, EEF evaluations commonly explore differential effects for FSM pupils. However, this study does not focus on individual pupils, and the theory of change does not indicate the percentage of FSM pupils at a school as a key influence, so we do not propose FSM subgroup analyses.

Treatment effects in the presence of non-compliance

Given the complex nature of the programme, there are various indicators that could be used to capture the dosage and fidelity (see Tables 10 and 11 below). Consequently, there are several potential approaches to defining compliance in *Aspire to STEM*. In our understanding, the first steps of programme participation (forming a partnership, conducting needs assessment and creating action plans) are mandatory for schools to be included in the programme, and therefore unlikely to have non-compliance among study schools.

Our suggested strategy to determine compliance²⁰, developed in discussions with STEM Learning, would be to create a continuous indicator primarily focusing on dosage, namely (i) the amount of spending of the budget (which can be accessed at school level)²¹; and (ii) the amount of teacher and leadership CPD, as these are the key activities of the programme. We would look to combine these measures if possible. To provide sufficient statistical power, we focus on compliance at *school* rather than *partnership* level.

Given imperfect compliance, we will re-run the analyses focusing on schools considered to be compliant with the treatment (and dropping the schools classified as non-compliant), to examine the complier intervention effects.

Missing data

The proposed approach relies on complete data. Using school-level data should minimise/negate missingness, although the proposal to match schools using longitudinal data brings further challenges in terms of inconsistencies. If schools closed/reopened, our suggestion is that previous results under old school names are used, rather than treating that

²⁰ Please See Implementation and process evaluation section below for definitions of compliance and dosage.

²¹ A budget indicator at school level would be weighted by the number of pupils, i.e. schools size, such as $X1 = N1 * Y / (N1 + N2 + N3)$, where X1 is the weighted budget, N1 is the school size, Y is the spending and N1-N3 are the sizes of other schools in the partnership.

information as missing. To aid with this task, we will use the Consistent Schools Database to ensure consistency over time with regard to schools data.²² One challenge is that the database is only available up to 2014, so matching schools IDs in the most recent years needs to be done manually. The other benefit of using TWANG is that the algorithm automatically generates dummy indicators for missingness, and thus attempts to balance on missingness as well (Ridgeway et al., 2017).

Effect size calculation

We will draw on the Hedges g adapted from Hedges (2007) as given below:

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$$

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean post-intervention difference between intervention groups adjusted for baseline characteristics (estimated in our case through the difference-in-difference coefficient) and $\sqrt{\sigma_S^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (variance) at baseline. The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchison and Styles, 2010). If we are able to use data from all TLIF-eligible schools, we will draw on the population variance, and if not, we will be drawing on the sample-adjusted variance. An alternative is offered by the What Works Clearinghouse Procedures Handbook, which discusses computing Hedges g of the difference between two groups on the baseline and post-intervention measures separately, and subtracting the baseline effect size from the post-intervention effect size (which would be done following weighting in this case).²³

Implementation and process evaluation

Following the EEF guidelines to process evaluation (Humphrey et al., 2015), we will conduct a process evaluation to better understand the work of this multi-component intervention. Since the initial proposal, the development of the IPE has been further informed by conversations with STEM Learning and the Intervention Delivery and Evaluation Analysis (IDEA) workshop we conducted.²⁴ The workshop was used to specify in more detail the mechanisms by which the intervention is expected to affect pupil and teacher outcomes, and it helped update Theory of Change for Aspire to STEM (see Figure 1).

The IPE will draw on the following types of data collection:

- Surveys with teachers;
- Surveys with headteacher;
- Surveys with educational leads;
- Surveys with parent groups/friends of school – if possible;
- School case studies involving phone interviews;
- Review of monitoring information collected by STEM Learning.

Given the quasi-experimental nature of the study, the IPE will include only the intervention schools (N=206), further excluding any schools that drop out from the programme after the

²² <https://www.closer.ac.uk/work-package-5-csd>

²³ https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf

²⁴ Goals of the workshop: 1. Agree the TIDieR framework for the intervention, 2. Discuss the intervention logic model, 3. Align measurement instruments for process evaluation to the intervention logic. Attendees: Ben Dunn, Andy Thirwell (STEM Learning), Julie Belanger, Sashka Dimova, Yulia Shenderovich (RAND Europe). Date of workshop: 9 April 2018

“quasi-randomisation” date because the IPE questions would not be relevant to them. We will consider targeted surveys or interviews with drop-out schools if the number is substantial.

The purpose of the process evaluation will be to address the following questions:

- What was the fidelity in the intervention schools?
- What was the dosage in the intervention schools?
- What factors and initial conditions appear to explain variation in the dosage and fidelity of implementation?
- What appear to be the necessary conditions for success of the intervention?
- What were the barriers to delivery?

Once the schools join the programme, they have access to STEM Learning network and resources: STEM Ambassador volunteers, STEM Club infrastructure and community groups and resource eLibrary. Next, as a key step of their participation, schools have to complete needs analyses and action plans for each school and agree on the activities for the partnership of between 3 and 10 schools. As described above, educational leads are consultants who guide schools through the intervention process, in particular completing their needs analyses and action plans. Once the action plan is agreed with the educational lead for each school in the partnership, as the logic model demonstrates, there are several types of activities and resources that schools can receive as part of the intervention:

- Bespoke CPD for STEM teachers (both on subject knowledge and subject pedagogy);
- Bespoke CPD for STEM leaders;
- Guidance on STEM careers for teacher and leaders ;
- Inspiration and enrichment activities.

We draw on the following definitions in our IPE (from Humphrey et al., 2015):

- **Fidelity/adherence** – the extent to which implementers adhere to the intended treatment model;
- **Dosage** – how much of the intended intervention has been delivered and/or received.

Following the EEF statistical analysis guidance, we define **compliance** as the extent to which the schools adhere to being in the treatment group.

The IPE needs to accommodate the bespoke nature of the intervention as the actual CPD events and other activities will vary depending on the needs and capacity of each partnership. Therefore, rather than focusing on the fidelity of individual trainings, for instance, we suggest to focus on fidelity to the overall model, e.g. the school partnership model.

As there is no single indicator for fidelity, we will rely on surveys with educational leads and headteachers as well as the interviews in case study schools to understand the extent to which the intervention was implemented as intended by the developers (see Table 10). Similarly, when examining the barriers and facilitators of delivery, we will aim to explore the functioning of the school partnership as part of the Aspire to STEM model. For instance, we will look at the process of agreement on programme activities within the partnership and the influence of the differences and similarities in the priorities of individual schools.

Table 10. Suggested fidelity indicators

Fidelity indicator	Data	Notes
Each school conducts needs assessment and action plan	<ul style="list-style-type: none"> • STEM Learning records • Educational leads surveys 	In principle, this is a mandatory component that all schools will most likely complete.
The schools are engaged in the partnership	<ul style="list-style-type: none"> • Educational leads surveys • Headteacher, STEM leaders and teacher surveys 	

Regarding dosage, given the number and diversity of potential intervention activities, there is not one straightforward measure of how much intervention schools receive. However, the evaluation can draw on the monitoring data collected by STEM Learning to paint a picture of the dosage of intervention activities (see Table 11).

Table 11. Suggested dosage indicators

Activity/resource	Data	Notes
Date of school needs assessment submission	Potentially measured by STEM Learning records	Reflecting the effective duration of the intervention
Access to STEM Learning network and resources: STEM Ambassador volunteers, STEM Club infrastructure and community groups and resource eLibrary	Not measured	It was discussed that access to resources would be too difficult to measure, especially as many online materials are available without a login.
Careers guidance and information for teachers and leaders	Number of career information events for teachers	Aggregated at school level by STEM Learning
Bespoke CPD for STEM teachers	Percentage of teachers attending CPD activities	Aggregated at school level by STEM Learning
Bespoke CPD for leaders	Percentage of school leaders attending CPD activities	Aggregated at school level by STEM Learning
Inspiration and Enrichment Activities	Number and type of community events	Aggregated at school level by STEM Learning
The schools remain in the partnership for the duration of the intervention (until March 2020)	<ul style="list-style-type: none"> • STEM Learning records • Educational leads surveys Headteacher, STEM leaders and teacher surveys	Assessed based on self-report by head-teachers and educational lead reports on whether all school continue to attend meetings and respond to communication.
Each school participates in at least some intervention activities (CPD, enrichment, etc.).	<ul style="list-style-type: none"> • STEM Learning records • Head teacher surveys 	
The partnership uses the available funding by the end of the project	<ul style="list-style-type: none"> • STEM Learning records 	

The intervention had started with some schools in Cohort 1 before the evaluation began, it is thus not possible to conduct a pre-intervention survey. However, in the initial survey for headteachers we will ask about non-STEM Learning practices regarding CPD and STEM engagement. Given the high turnover of school staff, we will not be aiming to follow up the same school staff. However, we can ask in the follow-up surveys if the respondent has completed the baseline survey.

For the purpose of IPE, four schools will be selected from each of the two cohorts as case study schools (eight schools total). In these schools, phone interviews will be conducted with several staff members (ideally, at least one teacher, a STEM subject leader and an SLT member) regarding the intervention. We suggest the following matrix for purposively selecting case studies, based on the baseline survey results and STEM Learning knowledge of the partnerships (Table 12). We will liaise with the STEM Learning team to define how similarity and difference of school priorities will be operationalised.

Table 12. IPE case study sampling frame

	Partnership with similar priorities	Partnership with diverse priorities
Previously existing partnership	1	1
New partnership	1	1

Finally, Table 13 provides an overview of the IPE data collection.

Table 13. Overview of the proposed IPE data collection

Data from	Mode	When	Who collects the data	Topics
Head teacher	Survey	1. October - November 2018	RAND to develop and analyse (STEM Learning to provide feedback and distribute the survey)	Survey 1: <ul style="list-style-type: none"> • The reasons for enrolment • Usual practice related to STEM CPD and STEM engagement • Previous participation in school partnerships or STEM Learning projects
		2. February –March 2020		Survey 2: <ul style="list-style-type: none"> • Experience/benefits of the partnership • Value-added of STEM ambassadors • Whether CPD/activities match their needs • Teacher retention • STEM Subject leadership • Programme cost (e.g. cover costs)/staff time – other than the 25K given to the partnership
Educational leads	Survey	1. October - November 2018	RAND (with STEM Learning feedback and contacts)	Survey 1: <ul style="list-style-type: none"> • Process of needs assessment/action plan creation • How similar or different were the needs of the schools in the partnership • Barriers and facilitators in the partnership planning process
		2. January/February 2020		Survey 2 <ul style="list-style-type: none"> • Follow-up on barriers and facilitators in the partnership cooperation once there have been CPD activities • Whether the action plan is being used as a tool for improvement
Primary teachers and secondary school STEM teachers	Survey	January/February 2020	RAND (with STEM Learning feedback and contacts)	<ul style="list-style-type: none"> • The experience/benefits of the partnership. • Whether CPD/activities match their needs. • Experience of working with STEM ambassadors. • Observed changes in student aspirations and future career awareness (both primary and secondary).
STEM subject leaders (faculty/departme	Survey	November 2019	RAND (with STEM Learning feedback and contacts)	<ul style="list-style-type: none"> • Perceived improved subject leadership • Active relationship with STEM employers

nt heads) and SLT)				<ul style="list-style-type: none"> Perceived change in engagement with parents, families and communities (a lot of schools now have a special person responsible for community engagement)
STEM teachers, head teachers, STEM subject leaders and SLT – in case study schools only	Interviews	December-January 2019	RAND (STEM Learning) to help identify partnership and interviewees for case studies)	<ul style="list-style-type: none"> Based on topics above and survey results
Educational lead monthly updates and the needs assessment/action plan – in case study schools only	Document review		RAND	<ul style="list-style-type: none"> Types of activities planned and actually implemented in the case study schools
Friends of school groups (usually parents)	Survey (originally discussed focus groups)		RAND	<ul style="list-style-type: none"> Community engagement in STEM learning, perceived benefits and challenges Conducted only if schools confirm to STEM Learning that it is feasible to distribute the surveys, and provide contacts

Cost evaluation

Cost data will be gathered through discussions with DfE, discussions with the delivery team at STEM Learning, as well as with questions included in the surveys and interviews to be used during the implementation process evaluation (see above). Questions will be targeted at assessing any pre-requisite costs (such as ensuring sufficient access to computers to access online resources from STEM Learning for example) and any direct and marginal costs directly attributable with schools' participation in the intervention (printing, materials, staff time, volunteer time, CPD costs, cover, etc.). We will use this information to estimate a cost per-pupil, following the EEF guidelines. The difficulty here is that 'business as usual' will likely include significant variation, so understanding AtS versus BAU in terms of costs will be difficult (and is likely to remain a limitation).

Ethics

The evaluation has been reviewed by RAND U.S. Human Subjects Protection Committee (HSPC) and approved on 12 June 2018. The evaluation includes a requirement to obtain informed consent only from school staff participating in the interviews. For the online surveys, since the burdens of participating in the survey are insignificant and sensitive topics are not involved, we will regard the completion of the survey as adequate evidence of consent (implicit consent).

Consent from other school staff or from pupils and their parents/guardians is not required as the intervention does not differ from standard practice in schools and the current evaluation is drawing on routinely collected school-level data.

None of the evaluation team has any conflicts of interest and all members of the study team have approved this protocol prior to publication.

Anticipated data sharing will be at school level and therefore not contain personal information. One exception is the teacher retention data, which will be received from DfE at teacher level but de-identified.

Data protection

Given the focus of the study outcome assessment on administrative data that does not contain personal intervention, the data protection mainly pertains to IPE surveys and interviews.

We will include a fair processing privacy notice in the preface to surveys and interviews, explaining why RAND Europe is collecting the data, how it will be stored and used. We will not be collecting participants' names in the surveys, so they will be anonymous (although school name will be collected to track response rates). In the survey initial screen, we will explicitly ask that participants do not provide any personal sensitive data in the survey, such as one's name or political persuasion. If such data are entered, they will be deleted.

As part of the interviews in case study schools, RAND will obtain personal data from school staff as a data controller. It is anticipated that basic staff information, such as names, required to organise the interviews, will be obtained from schools on the basis of legitimate interest.

The monitoring data shared by STEM Learning will be provided at school level and already anonymised, therefore not personally identifiable. The data will be shared via Syncplicity, a secure file sharing service, compliant with GDPR.

RAND Europe adopts good industry practices regarding the protection of personal data as part of its obligations as a Data Controller under the Data Protection Act 1998 and takes appropriate technical and organisational measures conformant with ISO 27001 to protect personal data. Individuals targeted by the study have the right to oppose, have access to, rectify, or remove personal or sensitive personal data held by RAND Europe.

Personnel

RECRUITMENT AND DELIVERY TEAM: STEM LEARNING

Project Officer: Andy Thirwell (STEM Learning)

Aspire to STEM Lead: Wayne Jarvis (STEM Learning)

Research and Evaluation Lead: Dr. Irina Kudenko

Research Assistant: Ben Dunn

EVALUATION TEAM: RAND

Project Leader: Dr. Alex Sutherland (RAND)

Project Manager: Dr. Yulia Shenderovich (RAND)

Core fieldwork and analysis team: Dr. Sashka Dimova (RAND), Jack Pollard (RAND)

Advisers: Dr. Kata Mihaly (RAND)

Risks

Risk	Assessment	Mitigation strategy
Recruitment failure	Likelihood: Low Impact: High	STEM Learning to recruit schools and be the main contact for schools. STEM Learning has already recruited the majority of the schools.
Attrition	Likelihood: Low Impact: Low	Given the reliance on administrative data, outcome data should be available for all schools, regardless of their participation in the project. Intention to Treat (ITT) analysis to be used.
Missing data	Likelihood: Moderate Impact: Moderate	Use of administrative data for pupils and teachers. Communication with DfE and experienced education researchers on questions regarding the data, for instance changes in school IDs .
Low implementation fidelity	Likelihood: Moderate Impact: Moderate	Process evaluation to monitor and document fidelity of implementation
Cross-contamination	Likelihood: Low Impact: High	Aspire to STEM is currently only delivered within TLIF. Information about other comparable STEM Learning and Institute of Physics programmes will be factored into the analysis.
Limited number of comparison schools	Likelihood: Low Impact: High	As part of weighting, all school in England with relevant data can be factored into the analyses.
Evaluation team members absence or turn-over	Likelihood: Moderate Impact: Low	All RAND staff have a three month notice period to allow sufficient time for handover. The team can be supplemented by researchers with experience in evaluation from the larger RAND Europe pool.
Low participation rates for IPE interviews	Likelihood: Moderate Impact: Moderate	Phone interviews to be offered at times convenient for school staff within a span of several weeks.
Low response rates for IPE surveys	Likelihood: Moderate Impact: Moderate	Surveys to be kept to 10-15 minutes long. Respondents given the opportunity to complete survey online on multiple occasions if required. Sufficient data collection window given with real-time monitoring of response rates to allow for reminders to be targeted.
Lack of coordination between RAND, STEM Learning, EEF	Likelihood: Moderate Impact: Moderate	Teams to attend initial meetings and agree on roles and responsibilities at the outset. Regular contact between senior team from each organisation.

Timeline (indicative)

- Timetable including specification of who is responsible for completing each task
- Please include specific dates or date intervals, even if only estimated (rather than, e.g., 'Autumn term').

Date	Activity	Staff leading responsible/
Feb 2018	Evaluation kick-off meetings	Bélanger (RAND)
February 2018	Intervention (needs assessment) begins in Cohort 1	Thirwell (STEM Learning); Jarvis (STEM Learning)
September 2018	Recruitment of schools completed	Thirwell (STEM Learning); Jarvis (STEM Learning)
September 2018	Intervention begins in Cohort 2	Thirwell (STEM Learning); Jarvis (STEM Learning)
October-November 2018	Initial surveys of headteachers and educational leads (cohorts 1 and 2)	Shenderovich (RAND)
December 2018-July 2019	Collation of weighting dataset for primary outcomes (pupil KS2 and KS4 outcomes)	Sutherland (RAND); Shenderovich (RAND)
November-December 2019	Teacher survey, Educational leads survey 2, STEM subject leader survey, Interviews in case study schools	Shenderovich (RAND)
February 2020	Head teacher survey 2	Shenderovich (RAND)
January 2021 – March 2021	KS2 and KS4 outcome analysis	
March 2020	Intervention completed in Cohorts 1 and 2	Thirwell (STEM Learning); Jarvis (STEM Learning)
April 2020	Sharing of STEM Learning monitoring data	Thirwell (STEM Learning); Jarvis (STEM Learning)
November 2021-June 2022	Teacher retention analysis	Sutherland (RAND)
March 2022-June 2022	KS5 progression analysis	Sutherland (RAND)
December 2022	Report on pupil and teacher retention outcomes	Sutherland (RAND)

Acknowledgements

We would like to thank colleagues from the Department for Education, in particular Pippa Gibson, for their input, as well as our quality assurance reviewers, Dr. Sam Sims and Dr. Marco Hafner. We are grateful to our former colleague Dr. Julie Belanger for her input in the early phase of the project and to Dr. Beth Ann Griffin for her input into finalising the analysis plan. We are also thankful to Guillermo Rodriguez-Guzmán (Education Endowment Foundation) who has provided valuable input into the study design.

References

- Abadie, A. and Gardeazabal, J. (2003). The economic costs of conflict. A case study of the Basque Country. *The American Economic Review*. 93(1), pp. 113-132.
- Abadie, A., Diamond, A., and Hainmueller, J. (2010). Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. *Journal of the American Statistical Association*. 105(490), pp. 493-505.
- Abadie, A., Diamond, A., and Hainmueller, J. (2015). Comparative politics and the synthetic control method. *American Journal of Political Science*. 59(2), pp. 495-510.
- Allen, R., and Sims, S. (2017a). Improving Science Teacher Retention: do National STEM Learning Network professional development courses keep science teachers in the classroom? The Wellcome Trust. <https://wellcome.ac.uk/sites/default/files/science-teacher-retention.pdf>
- Anders, J., Brown, C., Ehren, M., Greany, T. and Nelson, R. (2017). Evaluation of complex whole-school interventions: Methodological and practical considerations. A Report for the Education Endowment Foundation. UCL Institute of Education.
- Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology*. 2(1), pp. 23–44. <https://doi.org/10.1007/s11292-005-5126-x>
- Angrist, J. D. and Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*. 106(4), pp. 979–1014. <https://doi.org/10.2307/2937954>
- Angrist, J. D., & Pischke, J. S. (2008). Mostly harmless econometrics: An empiricist's companion. Princeton university press.
- Atteberry, A., Loeb, S. and Wyckoff, J. (2016). Teacher Churning: Reassignment Rates and Implications for Student Achievement. *Educational Evaluation and Policy Analysis*. 20(10), 1–28.
- Austin, P. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*. 46(3), pp. 399-424,
- Austin, P. (2011). Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharmaceutical Statistics*. 10, pp. 150-161
- Austin, P. (2014). A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*. 33, pp. 1057-1069.
- Austin, P., Grootendorst, P. and Anderson, G. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: A Monte Carlo study. *Statistics in Medicine*. 26(4), pp.734-753.
- Austin, P. C., & Small, D. S. (2014). The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in Medicine*. 33(24), 4306-4319.
- Banerjee, P. A. (2016). A systematic review of factors linked to poor academic performance of disadvantaged students in science and maths in schools. *Cogent Education*. 3(1), p. 1178441.

- Baser, O. (2006). Too much ado about propensity score models? Comparing methods of propensity score matching. *Value in Health*. 9(6), p. 2006
- Becker, S.O. and Caliendo, M. (2007) Sensitivity analysis for average treatment effects. *The Stata Journal* 7(1): 71–83.
- Belanger, J. and Broeks, M. (2016). Attracting and retaining teachers in Cambridgeshire: Working conditions and teacher flows from a School Workforce Census data perspective. RAND Corporation. Santa Monica, CA.
https://www.rand.org/pubs/research_reports/RR1850.html.
- Borenstein, M., Hedges, L., Higgins, J., and Rothstein, H. (2009). *Introduction to Meta-Analysis*. Sussex: John Wiley & Sons.
- Caliendo, M. and Kopeinig, S. (2005). Some practical guidance for the implementation of propensity score matching. No. 485, Discussion Papers of DIW Berlin, German Institute for Economic Research.
- Campbell, D.T. (1976). Focal local indicators for social program evaluation. *Social Indicators Research*. 3, pp. 237–256.
- Cefalu, M, Liu, S., Marti, C. (2015). Toolkit for Weighting and Analysis of Nonequivalent Groups: A tutorial on the Twang commands for Stata users. Available at:
<https://www.rand.org/pubs/tools/TL170.html>
- Dahabreh, I. J., Sheldrick, R. C., Paulus, J. K., Chung, M., Varvarigou, V., Jafri, H., Rassen, J.A., Trikalinos, T.A. and Kitsios, G. D. (2012). ‘Do observational studies using propensity score methods agree with randomized trials? A systematic comparison of studies on acute coronary syndromes’. *European Heart Journal*. 33(15), pp. 1893–1901.
- DfE (2016). Defining Achieving Excellence Areas Methodology guidance note. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/508392/Methodology_guidance_note_-_defining_achieving_excellence_areas.pdf
- DfE (2017). Opportunity Areas Selection Methodology. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650036/Opportunity_areas_selection_methodology.pdf
- Foster, D. (2018). Teacher recruitment and retention in England. House of Commons Library. Commons Briefing Papers CBP-7222.
<https://researchbriefings.parliament.uk/ResearchBriefing/Summary/CBP-7222>
- Funk, M. J., Westreich, D., Wiesen, C., Stürmer, T., Brookhart, M. A., & Davidian, M. (2011). Doubly robust estimation of causal effects. *American Journal of Epidemiology*. 173(7), pp. 761-767.
- Garet, M., Wayne, A., Stancavage, F., Taylor, J., Eaton, M., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolittle, F., Warner, E., (2011). Middle School Mathematics Professional Development Impact Study: Findings After the Second Year of Implementation (NCEE 2011-4024). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
- Garet, M. S., Heppen, J. B., Walters, K., Parkinson, J., Smith, T. M., Song, M., Garrett, R., Yang, R., Borman, G. D., & Wel, T. E. (2016). Focusing on mathematical knowledge: The impact of content-intensive teacher professional development (NCEE 2016-4010). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

- Gore, J., Lloyd, A., Smith, M., Bowe, J., Ellis, H. and Lubans, D. (2017). Effects of professional development on the quality of teaching: Results from a randomised controlled trial of Quality Teaching Rounds. *Teaching and Teacher Education*. 68, pp. 99-113.
- Griffin, B.A., Ridgeway, G., Morral, A.R., Burgette, L.F., Martin, C., Almirall, D., Ramchand, R., Jaycox, L. H. and McCaffrey, D.F. (2014). Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG) Website. RAND Corporation
<http://www.rand.org/statistics/twang>.
- Guo, S. and Fraser, M.W. (2010). *Propensity Score Analysis: Statistical Methods and Applications*. Advanced Quantitative Techniques in the Social Sciences Series (Number 12). London: SAGE.
- Harder, V. S., Stuart E. A., Anthony J.C. (2010). Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological Methods*. 15(3), p. 234
- Hedges, L. V. (2017). Challenges in Building Usable Knowledge in Education. *Journal of Research on Educational Effectiveness*. 11(1), pp. 1-21.
- Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016). Implementation and process evaluation (IPE) for interventions in education settings: An introductory handbook. Education Endowment Foundation.
- Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243-263.
- Jackson, L. (2017). Schools and class sizes in England & the UK. House of Commons Library. Commons Briefing Papers SN02625.
<https://researchbriefings.parliament.uk/ResearchBriefing/Summary/SN02625#fullreport>
- King, G and Nielsen, R. (2016) "Why Propensity Scores Should Not Be Used for Matching". Copy at <http://i.mp/2ovYGsW>
- King, G., Lucas, C. and Nielsen, R. (2017). The Balance-Sample Size Frontier in Matching Methods for Causal Inference. *American Journal of Political Science*. 61(2), pp. 473-489.
- Kirk, D.S. and Sampson, R.J. (2013). Juvenile arrest and collateral educational damage in the transition to adulthood. *Sociology of Education*. 86, pp. 336–362.
- Lee, B. K., Lessler, J., and Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine*. 29(3), pp. 337-346.
- Lee, D. and Lemieux, T. (2010). Regression Discontinuity Designs in Economics. *Journal of Economic Literature*. 48, pp. 281-355.
- Lockwood, J. R., and McCaffrey, D. F. (2016). Matching and Weighting With Functions of Error-Prone Covariates for Causal Inference. *Journal of the American Statistical Association*. 111(516), pp. 1831-1839.
- Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science*. 355(6325), pp. 584-585.
- Lunceford, J. K., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine*. 23(19), pp. 2937-2960.
- Maynard, R. A. and Dong, N. (2013). *PowerUp!*: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. *Journal of Research on Educational Effectiveness*. 6(1), pp. 24-67.

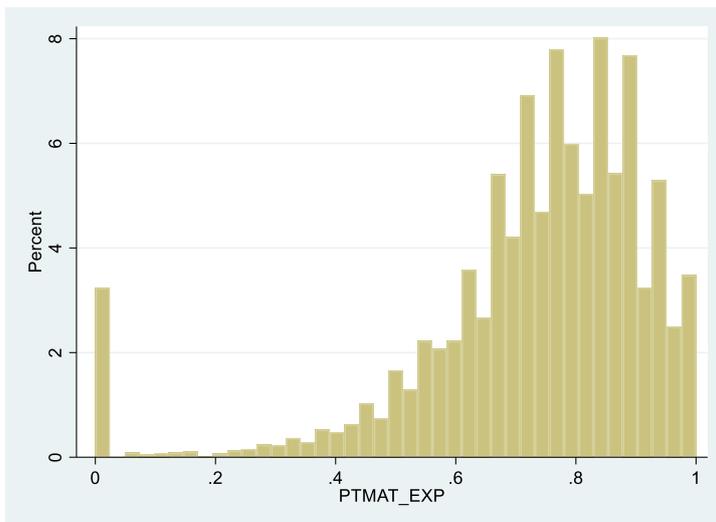
- McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., and Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in Medicine*. 32(19), pp. 3388-3414.
- McClelland, R. and Gault, S. (2017). The synthetic control method as a tool to understand state policy. Research report, The Urban Institute.
- McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*. 142(2), pp. 698-714.
- McKenzie, D. (2012). Beyond baseline and follow-up: The case for more T in experiments. *Journal of Development Economics*. 99(2), pp. 210-221.
- Nunes, T., Bryant, P., Strand, S., Hillier, J., Barros, R. and Miller-Friedmann, J. (2017). Review of SES and Science Learning in Formal Educational Settings. A Report Prepared for the EEF and the Royal Society.
- Opfer, D. (2016). Conditions and practices associated with teacher professional development and its impact on instruction in TALIS 2013. *OECD Education Working Papers*, No. 138. OECD Publishing, Paris.
- Pahlke, E., Hyde, J. S., and Allison, C. M. (2014). The effects of single-sex compared with coeducational schooling on students' performance and attitudes: A meta-analysis. *Psychological Bulletin*. 140(4), p. 1042.
- Parast, L., McCaffrey, D. F., Burgette, L. F., de la Guardia, F. H., Golinelli, D., Miles, J. N., and Griffin, B. A. (2017). Optimizing variance-bias trade-off in the TWANG package for estimation of propensity scores. *Health Services and Outcomes Research Methodology*. 17(3-4), pp. 175-197. <https://link.springer.com/article/10.1007/s10742-016-0168-2>
- Ridgeway, G., McCaffrey, D., Morral, A., Burgette, L., and Griffin, B. A. (2017). Toolkit for weighting and analysis of nonequivalent groups: A tutorial for the twang package. Available at: <https://cran.r-project.org/web/packages/twang/vignettes/twang.pdf>
- Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*. 70, pp. 41–55.
- Rosenbaum, P.R. (2002). *Observational Studies*. New York: Springer.
- Sims, S. and Fletcher-Wood, H. (2018) Characteristics of effective teacher professional development: what we know, what we don't, how we can find out. UCL Institute of Education. Available at: <https://improvingteaching.co.uk/wp-content/uploads/2018/09/Characteristics-of-Effective-Teacher-Professional-Development.pdf>
- Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Stat Sci*. 25(1), pp. 1-21.
- Stuart, E. A., Huskamp, H. A., Duckworth, K., Simmons, J., Song, Z., Chernew, M. E., & Barry, C. L. (2014). Using propensity scores in difference-in-differences models to estimate the effects of a policy change. *Health Services and Outcomes Research Methodology*. 14(4), pp. 166-182.
- Williamson, E., Morley, R., Lucas, R., and Carpenter, J. (2011). Propensity scores: From naive enthusiasm to intuitive understanding. *Statistical Methods in Medical Research*. 21 (3), pp. 273-283.
- Wing, C., Simon, K., & Bello-Gomez, R. A. (2018). Designing difference in difference studies: Best practices for public health policy research. *Annual Review of Public Health*. 39, pp. 453-469.

Woolridge, J. M. (2012). *Introductory econometrics: A modern approach*. Mason, OH: South-Western Cengage Learning.

Appendix 1.

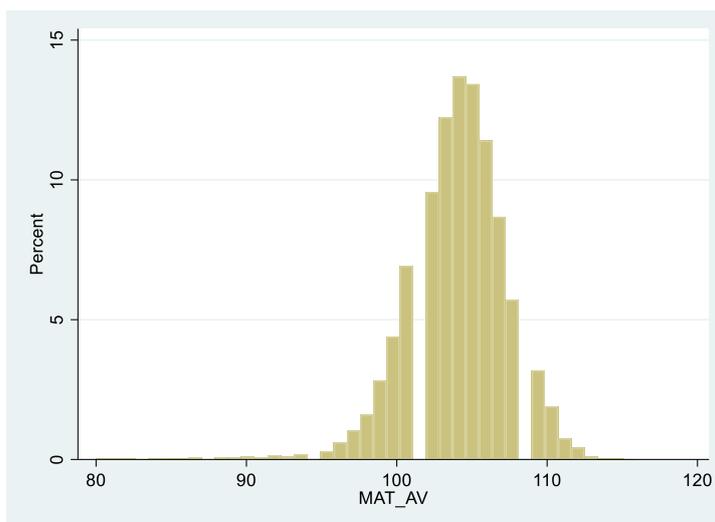
Options for KS2 Mathematics outcome

Based on our review of the school-level data, in 2016/17 the share of missing data are 7.0% for the “share of pupils achieving expected standards in maths” and 9.4% for “average scaled maths score”. This is because scores are not provided in the school-level public data for schools where pupils on average did not achieve the mark required for the lowest scaled score, whereas share of pupils achieve expected standards would be included as zero, thus reducing the share of missing data. The distributions and key characteristics for both variables are presented below.



Percent achieving expected standards

Variable	Obs	Mean	Std. Dev.	Min	Max
ExpStdPerc	15,168	73.1519	20.13622	0	100



Average scaled maths score

Variable	Obs	Mean	Std. Dev.	Min	Max
mat_av	14,774	103.9948	3.349966	80	116

ⁱ This template is designed to report non-experimental designs using data from the National Pupil Database. These guidelines are aligned with standards for reporting observational studies (GRACE, STARD and STROBE).
<https://www.graceprinciples.org/>; <http://www.stard-statement.org/>; <https://www.strobe-statement.org/index.php?id=strobe-home>