PSQM
Statistical Analysis Plan
Evaluator (institution): RAND Europe
Principal investigator(s): Elena Rosa Brown

Education Endowment Foundation

PROJECT TITLE	A randomised controlled trial of Primary Science Quality Mark
DEVELOPER (INSTITUTION)	PSQM, University of Hertfordshire
EVALUATOR (INSTITUTION)	RAND Europe
PRINCIPAL INVESTIGATOR(S)	Dr Alex Sutherland (20 July 2018 – 14 June 2019) Dr Emma Disley (15 June 2019 – 05 August 2019) Elena Rosa Brown (06 August 2019 – Present)
SAP AUTHOR(S)	Amelia Harshfield, Dr Sonia Ilie, Elena Rosa Brown, Miriam Broeks
TRIAL DESIGN	Two-arm stratified, cluster-randomised controlled trial
TRIAL TYPE	Efficacy
PUPIL AGE RANGE AND KEY STAGE	9 - 10, Year 5, KS2
NUMBER OF SCHOOLS	152
NUMBER OF PUPILS	4,097
PRIMARY OUTCOME MEASURE AND SOURCE	Primary outcome: science attainment in second year of trial. Source: To be confirmed. Likely an adapted version of the Thinking, Doing, Talking Science (TDTS) test if psychometric properties pass acceptability threshold (see updated protocol). Otherwise GL Assessment Progress Test in Science (PTS) will be used.
SECONDARY OUTCOME MEASURE AND SOURCE	Secondary outcomes: A) Pupil science attainment in first year of trial, B) pupil attitudes to science and science teaching. Source: A) same as primary measure, B) a newly-derived, piloted instrument, with items from TDTS and Trends in International Mathematics and Science Study (TIMMS)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	17/06/2020	*N/A*

Table of contents

Introduction 4
Design overview 5
Sample size calculations overview 6
Analysis 7
- Primary outcome analysis 7
- Secondary outcome analysis 9
- Subgroup analyses 9
- Additional analyses 10
- Longitudinal follow-up analyses 10
- Imbalance at baseline 10
- Missing data 11
- Compliance 12
- Intra-cluster correlations (ICCs) 13
- Effect size calculation 13

Introduction

Primary Science Quality Mark (PSQM) was initiated in 2008 at the University of Hertfordshire to raise the profile of science in primary schools in England and promote professional development in science teaching and leadership.[1,2] PSQM is a developmental accreditation programme aiming to improve science education in primary schools through providing teachers and school science leaders with a framework for self-assessment, reflection and development as well as relevant training.

PSQM is regularly delivered within geography-based hubs of schools to allow for schools to congregate for training days and discuss PSQM implementation. In this trial the hubs have a mean of 10 schools each and were used as a stratifying variable on account of schools working more closely together within their respective hub as opposed to outside it, and because of the single hub leader as below.

Each hub is supported by one experienced hub leader. Hub leaders have backgrounds such as Local Authority advisers, consultants, university lecturers and teachers who have achieved Primary Science Quality Marks in the past. Schools can work towards one of three Primary Science Quality Marks – PSQM, PSQM Gilt and PSQM Outreach. PSQM is for "schools which demonstrate how effective science leadership is *beginning* to have an impact on science teaching and learning across the school", whereas PSQM Gilt requires the demonstration of a "*sustained* impact", and PSQM Outreach is for schools that meet Gilt criteria and also impact science leadership and teaching in other schools.

Over the course of one academic year, PSQM involves a number of activities including staff training provided by the hub leader, developing science teaching in the school according to the PSQM framework, and online mentoring of subject leaders by hub leaders.

Awards are made to schools following an analysis of a series of documents that detail how the activities implemented during the intervention year have impacted on the science teaching and learning across the school and how the school meets the PSQM criteria. There are 13 PSQM criteria covering (1) primary school science leadership, (2) teaching (3) learning, and (4) wider opportunities, rather than the award itself being central, *the focus of the programme is on the process of self-assessment, reflection and development.*

All schools must complete the same self-evaluation and meet the same criteria, ensure that the subject leader (and another member of staff if possible) attend training, write and implement an action plan and submit common core documents. However, each school's action plan, implementation and final submission is relevant to its own context.

In the current trial, PSQM will be delivered in 76 primary schools, with another 76 schools assigned to the control arm. In the current evaluation, the programme will focus on the school's science subject leader and Year 5 teacher from each school (and a Key Stage (KS) 1 teacher, if the Year 5 teacher is the subject leader). More precisely, we will work with only one Year 5 teacher per school per year. In the event that the Year 5 teacher changes across the two years of the trial, we will aim to identify the reason for why the change occurred and start working with the new Year 5 teacher.

The PSQM programme is led by The University of Hertfordshire and will be independently evaluated by RAND Europe. The study is funded by the Education Endowment Foundation (EEF) and Wellcome Trust.

[1] http://www.psqm.org.uk/what-is-psqm
http://www.psqm.org.uk/about-us
[2] http://www.psqm.org.uk/__data/assets/pdf_file/0010/123130/Primary-Science-May-2016-PSQM-update.pdf

Design overview

The impact evaluation is designed to investigate the following research hypotheses:

Hypothesis 1: Year 5 pupils in randomly allocated primary schools participating in PSQM (intervention schools) will have higher levels of science attainment than the pupils in the comparison schools one year following the end of PSQM implementation, 2020/21 (Summer 2021; primary outcome).

Hypothesis 2: Year 5 pupils in primary schools participating in PSQM (intervention schools) will be report higher levels of enjoying science than the pupils in the comparison schools in 2020/21 (Summer 2021; secondary outcome).

Trial design, including number of arms		Two-arm stratified, cluster-randomised controlled trial, randomised at the school level
Unit of randomisation		School
Stratification variables (if applicable)		Region (hub) Note that the original protocol foresaw stratification on school size (operationalised as single/multi-entry schools); this was not implemented at randomisation stage due to lack of data.
Primary outcome	Variable	Science attainment, measured at the end of Study Year 2 (Summer 2021).
	measure (instrument, scale, source)	To be confirmed. Likely an adapted version of the Thinking, Doing, Talking Science (TDTS) test if psychometric properties pass acceptability threshold (the instrument is currently being re-designed by a team at York University, see updated protocol). Otherwise the GL Assessment Progress Test in Science (PTS) will be used.
Secondary outcome(s)	variable(s)	Pupil attitudes to science and science teaching (Study Year 2 (Summer 2021))
	measure(s) (instrument, scale, source)	Attainment: as per the primary outcome measure Attitudes: a newly-derived, piloted instrument, with items from TDTS and Trends in International Mathematics and Science Study (TIMMS)
Baseline for primary outcome	**Variable**	A standardised and pooled score: Mathematics, reading and writing (grammar/punctuation/spelling) at KS1 (procedure outlined in Baseline Measures)
	measure (instrument, scale, source)	National Pupil Database (NPD)
Baseline for secondary outcome	**Variable**	Mathematics, reading and writing (grammar/punctuation/spelling) at KS1
	measure (instrument, scale, source)	National Pupil Database (NPD)

The PSQM evaluation is a two-group parallel, stratified, cluster-randomised trial, with school as the unit of randomisation. To ensure comparability of schools in the intervention arm and the control arm

('exchangeability'),[3] we will randomise schools within hubs, which will serve to balance the study arms on geographical location and, therefore, any regional differences.

During recruitment period (2018-19 academic year), schools are asked to nominate one Year 5 teacher (in case there are multiple Year 5 classes) to participate in PSQM. The class of this teacher is considered the focal class for the evaluation, assessed in the summer 2021, a year after the implementation of PSQM. If by 2020-21 the teacher has moved, and there is more than one Year 5 class, a class will be randomly selected for assessment. We will also attempt to collect data on teacher attrition wherever possible; this information can then be integrated into an exploratory analysis that will test the hypothesis of whether PSQM is more effective when teachers are retained for the duration of the trial. We will then explore these results in relation to the logic model for PSQM.

To minimise burden on pupils and schools, the evaluation relies on administrative data at baseline, with schools providing pupil identifiers, which will be linked to the National Pupil Database (NPD). After schools have been recruited and the pupil and teacher information collected, the Evaluation Team will randomise schools to one of two arms: intervention or control.

Intervention schools will not be charged to take part in the PSQM programme and will receive a payment of £1,500 towards teaching cover and £120 towards travel costs. Control schools will not be allowed to participate in PSQM while the study is running but they will receive a payment of £1,500 on completion of the trial.

Sample size calculations overview

Power and minimum detectable effect size (MDES) calculations were performed using the PowerUp tools for main effects (Dong & Maynard, 2013) and moderators (Spybrook, Kelcey, & Dong, 2016; Dong, et al., 2017). Based on the EEF guidelines (EEF, 2018) and a recent evaluation working with science outcomes in this age group (Kitmitto 2018), the amount of variation explained by covariates for 140 schools with an average of 25 pupils each, is assumed to be 0.40 (equivalent to correlation of 0.63) for level 1 (pupils) and 0.00 for level 2 (schools). The efficacy evaluation of Thinking, Doing, Talking Science (TDTS), which used the same primary outcome (Hanley et al., 2015) reported intracluster correlation (ICC) of 0.15 in the analyses. With one class per school included in the evaluation, we assumed an average cluster size of 25 pupils. We also assume an alpha of 5% and an intended 80% power to detect effects. We use two-level clustered designs, assuming a continuous, normally distributed (Gaussian) outcome. Stratification variables, including the geographical hubs to which the schools belong, will be included in the final analysis model. We have not accounted separately for stratification in the power calculations.

Using the parameters above and with equal allocation to intervention and control the MDES in the protocol calculation of 140 schools is 0.197 (see Table 1). When calculating the MDES for the actual randomised data, the MDES comes to 0.188. We believe it would be important to power 0.2 even though this is an efficacy trial because the nature of the intervention is likely to result in comparatively smaller effect sizes.

We focus on a moderator effect defined as a statistical interaction of intervention and the moderator variable, in this case a binary variable identifying FSM eligibility at the pupil level.[4] Based on average number of free school meals (FSM) pupils in UK primary schools – 14% in 2018 - we assumed 4 FSM pupils per class for the protocol calculation, acknowledging that the PSQM recruitment for the trial focused on high-FSM area.

[3] For more information please see Oakes, J. M. (2013) Effect identification in comparative effectiveness research. The Journal for Electronic Health Data and Methods, 1(1):1004. doi: 10.13063/2327-9214.1004

[4] Please see the Statistical Analysis Guidance.

At randomisation, using baseline data, the actual mean number of FSM pupils per class comes to 765. Using the same assumptions as the main analysis, the sub-group MDES at protocol would have been 0.197; at randomisation the MDES is 0.161[5].

Table 1 presents the full breakdown of the power calculations both at protocol stage and at randomisation stage.

Table 1: power calculation based on protocol stage and post-randomisation stage

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM[b]
Minimum Detectable Effect Size (MDES)		0.197	0.251	0.188	0.161
Pre-test/ post-test correlations	level 1 (pupil)	0.63	0.63	0.63	0.63
	level 2 (class)	N/A	N/A	N/A	N/A
	level 3 (school)	0	0	0	0
Intracluster correlations (ICCs)	level 2 (class)	N/A	N/A	N/A	N/A
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha[6]		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		Two	Two	Two	Two
Average cluster size		25	4[a]	27	27[c] (5FSM)
Number of schools[7]	Intervention	70	70	76	76
	Control	70	70	76	76
	Total	140	140	152	152
Number of pupils	Intervention	1,750	1,750	2,071	2,071 (405[d] FSM)
	Control	1,750	1,750	2,027	2,027 (360 FSM)
	Total	3,500	3,500	4,098	4,098 (765 FSM)

[a] This is the average number of FSM pupils per class assumed.
[b] Analysis for sub-group run as moderator (interaction term). If sub-group analysis done on sub-sample, MDES for same parameters = 0.230.
[c] Analysis for sub-group run as interaction, therefore group size remains the same as in overall analysis
[d] One single pupil is yet to be confirmed if eligible for FSM. Therefore this number may become 406.

Analysis[8]

Primary outcome analysis

The aim of PSQM is to improve science education in primary schools. Therefore, the primary outcome will be pupil's science attainment. If the Thinking, Doing, Talking Science (TDTS) test by Hanley 2015

[5] Using the same parameters but applying the sub-sample calculation as per protocol, the FSM sub-group MDES at randomisation would be 0.230.

[8] Please see the Statistical Analysis Guidance.

achieves psychometric properties that pass acceptability thresholds, science attainment will be measured using an adapted version of it. This test was originally compiled from questions developed by Terry Russell and Linda McGuigan for an unrelated Randomised Controlled Trial (RCT) funded by the Wellcome Trust and covers a range of topics in biology, chemistry and physics. It included process/inquiry-based, concept-based; and open-ended conceptually-based questions. The reason for the adaptation is to align the test with the latest changes in the national curriculum and reflect the focus on science inquiry.

At the time of writing the proposal for this trial, the EEF commissioned a piece of work to an external team to compare the TDTS instrument to the current national curriculum to assess to what extent it measured up to the current standards and procedures. Their conclusion was that there is a poor match between the current version of the TDTS test and that of the current national curriculum. Thus, while the preferred primary outcome measure continues to be the TDTS test, this will rely in part on any adaptations to the test being externally tested (which are underway at the time of writing this statistical analysis plan), and for these to show good psychometric properties. In the case that there are poor psychometric properties, the evaluation team will use another measure for the primary outcome of this trial. This will be the GL Assessment Progress Test in Science (PTS).

At outcome, the chosen test will be administered in paper format and marked by a third-party, the National Foundation for Educational Research (NFER) to all Year 5 pupils registered in participating schools. This approach allows for blinding to allocation, as we can supply a list of schools to the assessors without revealing allocation. The duration of the test is expected to be 45 minutes.

The results of Year 5 pupils in the summer of the academic year 2020/21 (Cohort B) will be used as the primary outcome. The expectation is that given the nature of PSQM as a whole school intervention, intervention effects will take time to become visible. Using Cohort B outcomes as the primary outcome measure is therefore consistent with PSQM's theory of change. .[9] .

NFER will verify that the non-response rate for the both outcome data collection points is below 5%. If this is not the case, the analysis will follow the procedures set out in the 'Missing Data' section. Final data will then be supplied to RAND Europe for analysis. Schools will not be told in advance what the test is (i.e. the name of the test) but will be informed about the general areas it covers.

The primary outcome is change in science attainment, as measured by pupil-level science test (for Cohort B, standardised with mean 0 and standard deviation 1, and then pooled). We will use a two-level multilevel model to account for clustering of pupils in schools. Multilevel approaches assume that the schools in the study are a random sample of all schools and the multilevel modelling framework can flexibly handle complex variation within/between schools.[10]

The main analysis consists of the model for outcomes of pupils nested in schools, which is:

$$Y_{ij} = \beta_0 + \mathrm{PSQM}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \qquad (1)$$

where Y_{ij} is the science achievement of student *i* in school j (the primary outcome measure) PSQM_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here the stratifying variable of geographical location (hub) used for randomisation; X_{ij} represents characteristics at pupil level (pupil *i* in school j), specifically standardised baseline pupils scores (KS1 mathematics, reading and writing), standardised and pooled as for the outcome measure); $u_j$are referred to as school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and $e_{ij}$are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$). In relation to X_{ij}, this assumes that the KS1 measures are not too highly correlated with each other. If the pair-wise correlation between measures exceeds r=.7 (i.e. the

[9] Initial trial plans were to use the results of pupils in summer term of academic year 2019/2020 (Cohort A) as a secondary outcome. COVID-19 -related lockdown and social distancing measures resulted in the cancelation of this testing round.

[10] Snijders, Tom A.B., and Bosker, Roel J. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling, second edition. London etc.: Sage Publishers, 2012.

shared explained variance is over 50%), then only one measure will be included (KS1 maths score, treated in the same manner as above).

Equation (1) is known as a 'random intercepts' model because $\beta_{0j} = \beta_0 + u_j$ is interpreted as the school-specific intercept for school j and $\beta_{0j} \sim i.i.d\ N(\beta_0, \sigma_u^2)$ is random (as in it can take any value). The total residual variance can be partitioned into two components: the between-school variance $\sigma_u^2$and the within-school variance σ_e^2.

Our target parameter (i.e. the focal result of the trial) τ is the average effect of the intervention on pupil outcomes compared to control schools. The τ coefficient refers to the relationship between PSQM allocation and the outcome for Cohort B represents the main result of the trial. All analyses will be performed in Stata, versions 15.1 onwards.

The outcome analysis will be on an intention-to-treat (ITT) basis. This method compares outcome means for the treatment and comparison groups, and subjects are analysed according to their randomised group allocation. The ITT approach is inherently conservative as it captures the averaged effect of offering the intervention, regardless of whether or not the participants comply with the assignment.

Problems of dropout/non-attendance may be an issue for this trial depending on how motivated school staff are. The main concern is that new teachers come in or that schools and/or teachers and/or subject leaders drop out at some point during the trial. These risks are mitigated by this being a whole school intervention.

Secondary outcome analysis

There are two secondary outcomes for the PSQM trial:

The first secondary outcome is attitudes to science and science skills among Year 5 pupils. The attitudinal measure at post-test will also be administered by NFER using paper forms. The attitudinal measures will be compiled in machine-readable forms, to allow scanning, data entry and scoring by RAND Europe. Enjoyment of science, confidence in science and engaging teaching in science will be measured using the 'enjoyment of science' subscale adapted from the Trends in International Mathematics and Science Study (TIMMS) Grade 4 surveys from TIMMS 2015.

To ensure that this adapted attitudes to science measure is valid, a small validation pilot will be conducted prior to outcome measure collection in trial schools. This pilot will use a sample of Year 5 pupils (in non-trial schools) to test the measurement structure of the variable (in exploratory factor analysis). The pilot will yield a final measure to be used in the trial. This will generate a continuous score that will be then entered into Equation (2) for the purposes of analysis.

$$Y_{ij} = \beta_0 + \mathrm{PSQM}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \qquad (2)$$

where Y_{ij} is the science attitude of student *i* in school j (as captured with the science attitude measure above); PSQM_j is a binary indicator of the school assignment to intervention [1] or control [0]; Z_j are school-level characteristics, here the stratifying variable of geographical location (hub) used for randomisation; X_{ij} represents characteristics at pupil level (pupil *i* in school j), specifically standardised baseline pupils scores (KS1 mathematics, reading and writing), standardised and pooled as for the outcome measure); $u_j$are referred to as school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and $e_{ij}$are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$).

The unit of analysis for the secondary outcome analysis will be the pupil.

Subgroup analyses

With 152 schools (as at randomisation) and an interaction term approach to undertaking the sub-group analysis, the study should be powered for meaningful sub-group analysis, such as stratifying by

FSM pupils. This depends, however, on the level of attrition from the trial by the final analysis stage, and therefore it may be that the study will be under-powered for sub-group analyses at that stage.

We will report mean outcomes by sub-categories of FSM as a basic descriptive step. As an exploratory analysis we will do sub-group analyses for FSM, acknowledging that this analysis may be underpowered. As an exploratory modelling approach, EverFSM will be incorporated into the multilevel model as a binary variable [1] if EverFSM, [0] otherwise. The EverFSM indicator will then be interacted with treatment allocation to assess the conditional impact of PSQM on FSM pupils.

We will conduct the FSM sub-group analysis for the primary outcome only (i.e. using Cohort B data), using the model in Equation (3) below:

$$Y_{ij} = \beta_0 + (\mathrm{PSQM}_j \times FSM_i)\tau_s + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij} \quad (3)$$

where Y_{ij} is the achievement of student *i* in school j; $(\mathrm{PSQM}_j \times FSM_i)$ is a binary indicator of the school assignment to intervention [1] or control [0] interacted with the pupil-level binary eligibility for FSM variable; Z_j are school-level characteristics, here the stratifying variable of geographical location (hub) used for randomisation; X_{ij} represents characteristics at pupil level (pupil *i* in school j), specifically pupils' standardised baseline scores (KS1 mathematics, reading and writing), standardised and pooled as for the primary outcome measure analysis); $u_j$are referred to as school-level residuals ($u_j \sim i.i.d\ N(0, \sigma_u^2)$) and $e_{ij}$are individual-level residuals ($e_{ij} \sim i.i.d\ N(0, \sigma_e^2)$).The sum of the coefficient for the treatment and the coefficient for the interaction of the treatment variable with FSM will represent the result for the sub-group analysis.

A second set of sub-group (moderator) analyses will be carried out, again for Cohort B using gender (operationalised here as a binary variable, coded 1 if female, 0 if male) as the moderator. Equation 3 will be applied in the same manner as above, replacing the FSM indicator with a binary gender indicator.

As these analyses are exploratory and potentially underpowered, we would report point estimates and confidence intervals transformed into effect sizes but would not report significance tests/p-values.

Additional analyses

As mentioned previously we will carry out an exploratory analysis in relation to teacher retention for the duration of the trial. This will be undertaken in a two-stage least squares approach, for the main outcome of the trial only. This is outlined in full in the Compliance Analysis section below.

Longitudinal follow-up analyses[11]

No longitudinal follow-up analyses are planned following this trial. However, we will submit an anonymised version (i.e. with no personally identifiable data, nor sensitive data) of selected results following trial completion to the EEF Archive for the benefit of any potential future work in this area.

Imbalance at baseline

We have taken an active approach to address imbalance by stratifying the randomisation. A well-conducted randomisation will, in expectation, yield groups that are equivalent at baseline.[12] Because schools are randomly allocated to the control and intervention conditions, any imbalance at baseline will have occurred by chance. To check for, and monitor, imbalance at baseline in the realised randomisation, analyses will be conducted at the school and pupil level. At the school level, the analysis will look at the following variables, by means of cross-tabulations and histograms that assess the distribution of each characteristic within control and treatment groups aggregated from the pupil data in the study sample (rather than publicly available school-level statistics):

[11] Please see the longitudinal analysis guidance.

[12] Glennerster, R. and Takavarasha, K. (2013) Running randomized evaluations: a practical guide. London: Princeton University Press.

- Type of school (multi vs single entry[13]).
- Proportion of pupils eligible for FSM.
- Proportion of pupils speaking English as an additional language (EAL).
- School-level average KS1 scores (for sample pupils).

At the pupil level, the initial balance will be assessed for the following characteristics:

- Eligibility for FSM (EverFSM).
- Gender.
- KS1 attainment (expressed as a standardised mean difference).

Statistical significance tests will not be carried out to assess the balance, as their premise does not hold in randomised control trials[14] (i.e. given appropriate randomisation procedures were followed, any differences between control and treatment groups at baseline will be by definition due to chance, and classical statistical testing is therefore unnecessary). Instead, tables of the means (and standard deviation, where appropriate) for each characteristic will be presented along with distributions. Where imbalance is found, and in relation to covariates that are found to be predictive of the outcome, the magnitude of any differences will be explored[15] and a decision made as to whether they require inclusion in the analysis.[16]. If the difference is higher than what convention dictates[13], then the respective covariates will be entered into the final analysis models.

Missing data

Missing data can arise from item non-response or attrition of participants at school, teacher and pupil levels. We will first determine the proportion of missing data in the trial. Our use of administrative data for pupil baseline data should reduce missingness arising. Below we set out our missing data strategy.

We will explore attrition across trial arms as a basic step to assess bias.[17] We will provide cross-tabulations of the proportions of missing values on all baseline characteristics (as detailed in the previous section, at both pupil and school level), as well as on the primary outcome measures.

To assess whether there are systematic differences between those who drop out and those who do not – and thus whether these factors should be included in analysis – we will model missingness at follow-up as a function of baseline covariates, including treatment. The analysis model for this approach will mirror the multilevel level model given above (pupils clustered in classes), but the outcome will be a binary variable identifying missingness (yes/no).

For less than 5% missingness overall, a complete-case analysis might suffice (i.e. assuming data are Missing Completely at Random (MCAR)), but our default will be to check results using approaches that account for missingness but that rely on the weaker Missing at Random (MAR) assumption. Our preference is to use Full-Information Maximum Likelihood (FIML) over multiple-imputation (MI) because FIML can be estimated in a single model and simulation studies show that it can reduce bias

[13] Please note that where a school is mixed in its approach to multi vs single entry, we will use the baseline status of school for Year 5.
[14] http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data
[15] There is a convention in some disciplines that a 10pp (or larger) difference in treatment and control means at baseline constitutes 'imbalance' is thus justification for including those measures in sensitivity analyses, but there are counter-arguments to this idea (see Roberts,C. and Torgerson, D. (1999) 'Baseline imbalance in randomised controlled trials', BMJ, 319:185; but also see de Boer et al. (2015) 'Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate', International Journal of Behavioral Nutrition and Physical Activity, 12:4).
[16] Senn, S. (1994) 'Testing for baseline balance in clinical trials', Statistics in Medicine, 13: 1715-1726.
[17] Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., & Sterne, J. A. C. (2011) The Cochrane Collaboration's tool for assessing risk of bias in randomised trials, British Medical Journal, 343(d5928).

as well as MI (for a discussion of FIML vs MI see Allison, 2012[18]). (For missingness on outcome variables only standard statistical packages such as Stata use ML for estimating parameters so FIML would not be necessary.[18]

Compliance

The main framework of analysis for this trial is Intention-to-Treat (ITT). This means that we will use the original school allocations in the final analyses, regardless of the level of compliance or implementation fidelity.

However, we will also be able to explore the effect of the intervention on schools that were allocated to the intervention group and also implemented the intervention.

In collaboration with the delivery team and drawing on the logic model for the intervention we have derived a binary compliance measure. We have defined "compliance" as the fulfilment of a set of minimum criteria which determine whether a school has effectively participated in PSQM. This is a binary measure, indicating whether a school is compliant or not. It involves scoring separate elements of the implementation using binary variables to capture whether an identified dimension is met by the school. Missing data on any measure will be scored as zero. Schools will not have sight of the compliance measure checklist.

Table 2 specifies the three dimensions that will be considered for the compliance metric. These are training attendance – distinguishing between science subject leader and Year 5 teacher attendance – and PSQM task completion. These dimensions are equality important and will receive equal weighting when arriving at the final compliance score per school.

The evaluation team will receive training attendance logs collected by PSQM to assess on a school-by-school basis whether the science subject leader and Year 5 teacher attended the training sessions. There are four sessions in total. It is mandatory for science subject leaders to attend all four sessions. In the case of Year 5 teachers, they are encouraged to join the training, but lack of attendance was considered to have less of a negative effect on programme impact by the delivery team. Therefore, upon discussion with the delivery team it was agreed that a school would be considered compliant with regards to teacher training attendance if the Year 5 teacher attended at least two training sessions.

In addition, the completion of PSQM tasks will also be considered. For this, the evaluation team will gain access to PSQM's Virtual Learning Environment (VLE) platform to check whether each required task is completed by each school. This will be evidenced by having uploaded relevant documentation for each PSQM task to the VLE. There are ten tasks schools are expected to complete and upload. Therefore, a school will be considered compliant if they have submitted all ten tasks in VLE.

Table 2: Dimensions contributing to the compliance metric

Compliance dimension	Variable (options)	Data source
Science subject leader attendance at training sessions -	Did the science subject leader attend all four training sessions? Yes/ No	Attendance logs from PSQM
Year 5 teacher attendance at training sessions	Did the Year 5 teacher attend at least two training sessions? Yes/ No	Attendance logs from PSQM
School task completion	Did the school complete all ten tasks? Yes/No	School task completion logs (looking into submission of common core documents in

[18] Allison, P. D., (2012) Handling Missing Data by Maximum Likelihood. Haverford, PA: Statistical Horizons. Retrieved from: https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf

VLE)

For a school to be deemed compliant, **all** three compliance criteria must be fulfilled. In other words, any school that only meets two of the three criteria will not be deemed fully compliant. We will then check how many of the schools have met compliance, but not use this binary variable (complied/not complied) in an analysis as this would no longer respect the randomisation logic and therefore resulting estimates are likely to be biased.

Instead, in the event of imperfect compliance (i.e. schools in the non-complier group, as defined above), we will use the compliance score derived above in an instrumental variable approach (via a two stage least squares regression, 2SLS) to derive a compliance 'instrument' that is not endogenous and can be used to estimate the effect of compliance to treatment. The 2SLS approach includes a first stage where the compliance measure is predicted from the treatment allocation indicator, and then this is used in the second stage for the outcome analysis in the presence of compliance.

Intra-cluster correlations (ICCs)

The between-school variance (ICC) will be calculated in the first instance, using a model with no predictors, but accounting for the clustering of pupils in schools (the so-called empty model).

This will be calculated separately for Cohort A (the attainment secondary outcome analysis) and Cohort B (the primary outcome analysis) as well as for the attitudinal secondary outcomes.

Effect size calculation

Given the multi-level approach to analysis across all the above models, we will use the effect sizes for cluster-randomised trials given in the EEF evaluator guidance; specifically, we will use the effect size calculation adapted from Hedges[19]:

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

Where $(\bar{Y}_T - \bar{Y}_C)_{adjusted}$ is the mean difference between intervention groups adjusted for baseline characteristics and $\sqrt{\sigma_S^2 + \sigma_{error}^2}$ is an estimate of the population standard deviation (variance). In the multi-level models this variance will be the total variance (across both pupil and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The effect size (ES) therefore represents the proportion of the population standard deviation attributable to the intervention.[20] A 95% confidence interval for the ES, that takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the regressions estimated.

[19] Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

[20] Hutchison, D., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. Slough: NFER.