PROJECT TITLE	**Evaluation of Focus4TAPS**
DEVELOPER (INSTITUTION)	Bath Spa University
EVALUATOR (INSTITUTION)	National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR(S)	Dr Hayley Leonard
PROTOCOL AUTHOR(S)	Dr Hayley Leonard, Dr Georges Poquillon, Sehaj Bhatti, Rebecca Parker, Dr Liza Benny, Kate Wadsworth
TRIAL DESIGN	Two-arm clustered randomised controlled trial with random allocation at school level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	Year 5: ages 9 to 10; KS2
NUMBER OF SCHOOLS	240
NUMBER OF PUPILS	6,371 (25 pupils per school)
PRIMARY OUTCOME MEASURE AND SOURCE	Science attainment: mark achieved on science assessment test developed by University of York (Joshi et al., 2021) for the EEF with a continuous single scale (0-45)
SECONDARY OUTCOME MEASURE AND SOURCE	Attitudes to science: Bespoke survey questionnaire developed by Focus4TAPS efficacy trial evaluation team, and used in efficacy trial

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [original]		N/A

Table of contents

Introduction

This Statistical Analysis Plan (SAP) describes the planned impact evaluation analysis for an evaluation of a two-arm cluster randomised effectiveness trial of the Focus4TAPS intervention delivered to Year 5 pupils as part of their regular science lessons in the classroom.

Focus4TAPS is a Continuing Professional Development (CPD) programme that aims to build primary school teachers' confidence in teaching science, giving them a way to embed effective formative assessment in lessons, expected to lead to improved pupil learning outcomes in science.

The Focus4TAPS programme was developed by Bath Spa University (BSU). It is based on the Teacher Assessment in Primary Science (TAPS) project, which emphasises the active participation of pupils and the responsiveness of teachers as they utilise assessment for learning. Each school taking part in the evaluation nominated two teachers (one Year 5 science teacher and one science subject lead) to attend the training run by BSU. If both roles were held by the same teacher, only one teacher attended the training. Year 5 pupils in the classes of teachers who have had the training are the recipients of the programme. Only one Year 5 class from each school are included in the trial, although the teaching approach can be used with other classes who are not involved in the evaluation activities.

In the effectiveness trial, the delivery mode for the intervention has been adapted. to a train-the-trainer approach, adding an intermediary between the programme developer and teachers. To support scale-up, the training programme was adapted to be delivered by designated trainers rather than directly by the programme developers. In contrast, the efficacy trial used a direct approach where programme developers trained teachers directly. The effectiveness trial expands delivery to a larger sample of primary school children across a larger number of regions in England. More information regarding the evaluation of the Focus4TAPS effectiveness trial is available in the Focus4TAPS trial protocol (Leonard et al., 2024).

Design overview

The impact evaluation is designed as a two-arm cluster randomised effectiveness trial of the effect of Focus4TAPS on Year 5 science attainment. The unit of randomisation is at the school-level, and unit of analysis is at the pupil-level.

240 schools were recruited for the trial and randomised into the Focus4TAPS intervention group or the control group on 3 July 2024. BSU were responsible for recruiting schools for the trial and recruited 240 schools in total across the following regions: Newcastle, Sunderland, Sheffield, Liverpool, Bury, Nottingham, Derby, Wolverhampton, Suffolk, Essex, Southampton, South-East London and West London. The selection of regions was determined based on the availability of trainers and the concentration of schools within each area. We stratified randomisation by region and school level KS1 attainment. Schools provided the percentage of pupils

who had met the expected standard in KS1 reading and maths.[1] We calculated the average to create a single KS1 attainment measure for each school and the sample median of this average was used to classify schools as being either above or below the median attainment for the sample. Settings were equally likely to be allocated to the Focus4TAPS intervention or control groups within each region-KS1 attainment strata.

Schools assigned to the Focus4TAPS intervention group are offered the 3-day Focus4TAPS training during the 2024-25 academic year at a reduced rate of £150. Schools assigned to the control group will implement a teaching-as-usual approach to science teaching. Schools are provided with an additional financial incentive to participate in the trial. Focus4TAPS intervention and control schools will receive an additional incentive of £200 and £500, respectively, at the end of the trial for completing all the trial requirements. Control schools can also access the Focus4TAPS programme at a significantly reduced fee after the trial ends. These incentives are designed to mitigate the risk that schools choose not to participate after assignment to the control group, and to incentivise them to complete all trial requirements.

The primary outcome for the trial is science attainment using scores on a science attainment test developed by the University of York (Joshi et al., 2021) for the EEF, graded on a single continuous scale (0-45). The assessment test comprises fifteen questions (with 38 individual items) designed with balanced coverage across subject areas, topic areas, challenge levels, including coverage of the Working Scientifically subscale (Joshi et al., 2021). The questions on the assessment test involve a range of question formats, including multiple-choice, free-text, drawing, and graphical responses.

The primary baseline measures for the trial are Key Stage 1 (KS1) reading and KS1 mathematics classifications. These measures were requested from the National Pupil Database (NPD) in January 2025.

The secondary outcomes for the trial cover attitudes and beliefs to science. These will be measured using the questionnaire developed for the efficacy trial, consisting of seven subscales that define the secondary outcomes:

(1) Interest and enjoyment in science

(2) Confidence in science

(3) Perceptions of science teachers

(4) Self-regulation of learning in science

(5) Self-efficacy for Working Scientifically

(6) Working Scientifically beliefs

(7) Wider benefits of science.

[1] KS1 attainment data used in the trial are from the 2021/22 academic year, when the Year 5 pupils included in this trial completed their KS1 assessments, when assessment at the end of KS1 was a statutory requirement.

The attitudes and beliefs to science questionnaire is administered to pupils at both baseline and endline.

Schools completed administration of the baseline assessment by 14 October 2024 which will be used as the baseline measure for the secondary outcomes. In total, 223 schools (114 Focus4TAPS intervention[2], 109 control) completed the baseline assessment out of the 240 schools that were randomised. A total of 5,598 pupil surveys were completed at baseline. Table 1 summarises the trial design.

[2] One Focus4TAPS intervention school who completed the baseline assessment subsequently withdrew from the trial, so the total number of 'active' schools who completed the baseline assessment is 222.

Table 1: Trial design overview

Trial design, including number of arms		Two-arm cluster randomised effectiveness trial
Unit of randomisation		School
Stratification variables (if applicable)		Geographical region of the intervention training delivery School level prior KS1 maths and reading attainment
Primary outcome	variable	Science attainment
	measure (instrument, scale, source)	Mark achieved on a science attainment test developed by University of York (Joshi et al., 2021) for the EEF with a continuous single scale (0-45)
Secondary outcome(s)	variable(s)	Attitudes to science
	measure(s) (instrument, scale, source)	Questionnaire developed by Focus4TAPS efficacy trial evaluation team, and used in efficacy trial, assessing: • Interest/enjoyment in science • Confidence in science • Perceptions of science teachers and practices • Self-efficacy within Working Scientifically • Wider views concerning Working Scientifically • Wider views concerning science Scores for each factor are scaled from 1 ('Disagree a lot') to 4 ('Agree a lot').
Baseline for primary outcome	Variable(s)	Key Stage 1 reading and mathematics attainment
	measure (instrument, scale, source)	Key Stage 1 Reading classification: KS1_READ_OUTCOME Key Stage 1 Mathematics classification: KS1_MATH_OUTCOME (National Pupil Database)
Baseline for secondary outcome	variable	Attitudes and beliefs towards science
	measure (instrument, scale, source)	Baseline pupil survey data

Research questions

This impact evaluation aims to answer the following primary research question (IEQ):

> **IEQ1.** What is the impact of schools' participation in the Focus4TAPS programme on the science attainment of Year 5 pupils (as measured by the science attainment test), compared with teaching as usual?

The secondary research questions that the evaluation will also answer are:

> **IEQ2.** What is the impact of the Focus4TAPS programme on Year 5 pupils' attitudes towards science compared to teaching as usual?
>
> **IEQ3.** Is any impact (IEQ1 and IEQ2) moderated by eligibility for Free School Meals (FSM)?
>
> **IEQ4.** Is any impact (IEQ1 and IEQ2) moderated by any of the following characteristics?
>
> (a) gender;
>
> (b) exposure to science
>
> (c) KS1 maths category
>
> (d) pupils' pre-intervention interest in science
>
> **IEQ5.** What is the Complier Average Causal Effect (CACE) of the Focus4TAPS programme on the science attainment of Year 5 pupils (as measured by the science attainment test), compared with teaching as usual?
>
> **IEQ6.** What is the per-pupil cost of Focus4TAPS?

Randomisation

A total of 260 schools registered interest in participating in the trial as they returned Memorandums of Understanding, of which 20 schools withdrew or did not upload pupil data (or provide assurances that pupil data would be uploaded in the future) in time for randomisation. This resulted in a total of 240 schools being included for randomisation (assuming n = 6,000 pupils)[3].

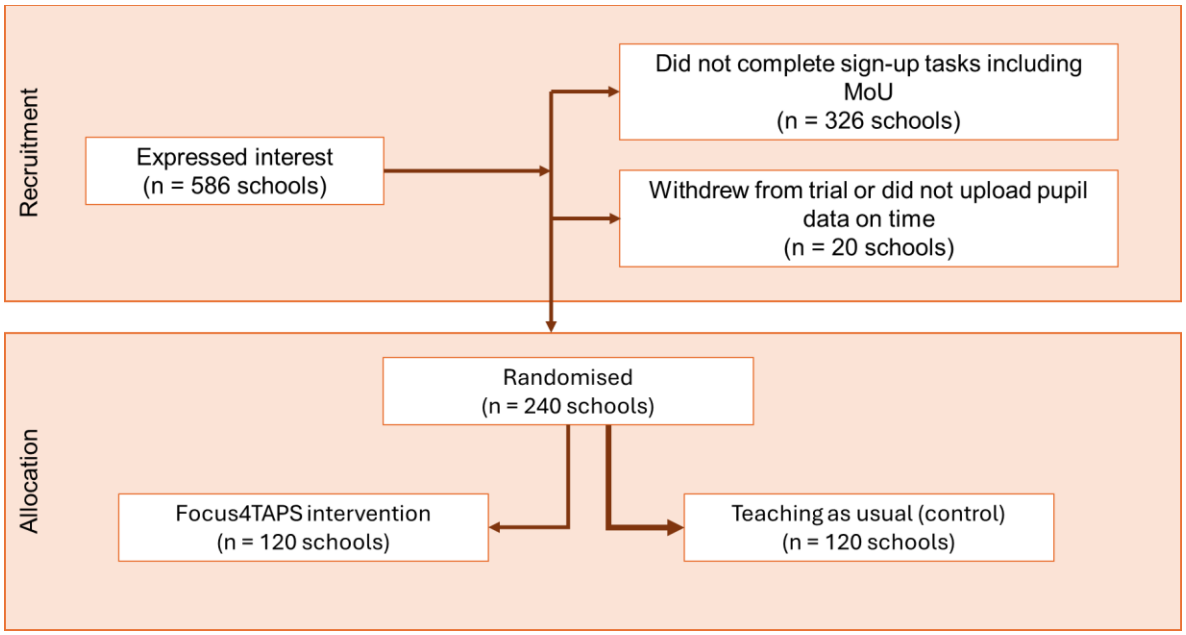
Randomisation of schools was carried out using Stata 17 by a member of the evaluation team on 03 July 2024.[4] Randomisation was carried out blind to school identity as meaningful identifiers such as school names were removed from the list of schools to be randomised and linked back in after randomisation was completed.

[3] Randomisation took place before all schools could upload pupil data, therefore the number of pupils at randomisation reflects pupil uploads received by the evaluation team at the time of writing. The 6,000 pupils available at randomisation were from 227 schools. Note that those pupil uploads were made in the previous school year (2023-2024), thus the actual number of pupils changed as for example some pupils moved to different schools. Pupil participation details were finalised after the new academic year started in Sep/Oct 2024, after most schools completed uploads of data.

[4] Stata syntax is included in Appendix A.

Randomisation was carried out using the *randtreat* command, using the misfits(global) option to allocate remaining schools when school numbers were uneven within strata. Randomisation was stratified to ensure that schools from the same region and with similar levels of mean KS1 maths and reading attainment were allocated evenly between the Focus4TAPS intervention and control groups. Figure 1 below summarises the flow of schools from recruitment to randomisation to intervention allocation.

Figure 1: Flow diagram showing school participation over trial stages



Totally, 240 schools were randomised simultaneously, with 120 schools allocated to the Focus4TAPS intervention group and 120 schools allocated to the control group. Table 2 shows intervention status within each stratum.[5] The number of schools varied across strata, with more schools in some regions such as Sunderland.[6] The randomisation produced a roughly equal allocation of schools in each stratum, though the small number of schools within the combined strata of region and school KS1 attainment meant that numbers were slightly unequal within these.

Table 2: Randomisation allocation across strata

Strata	Control	Treatment	Total
Region			
Bury	9	9	18
Derby	8	8	16
Essex	8	8	16
Liverpool	7	8	15
Newcastle	17	16	33

[5] Appendix Table B1 summarises the randomisation allocation within each strata combination.
[6] Recruitment levels varied by region. Schools in previously 'cold' regions with low levels of Focus4TAPS exposure were targeted as a focus for the programme. However, there were higher levels of interest in participation from schools in regions with established school networks and awareness of the programme, such as Sunderland and Newcastle. Some regions (such as Southampton) also lost initially interested schools to another EEF trial, which was recruiting similar schools for a primary science effectiveness trial.

Nottingham	5	6	11
SE London	9	8	17
Sheffield	11	12	23
Southampton	7	7	14
Suffolk	9	9	18
Sunderland	12	12	24
W London	7	7	14
Wolverhampton	11	10	21
KS1 attainment			
Below median	52	54	106
Above median	55	54	109
Missing KS1 scores	13	12	25
Total	**120**	**120**	**240**

The randomisation allocation was recorded in Excel and communicated to the delivery team on 3 July 2024 to facilitate onboarding meetings where schools were separately introduced to the programme in each region by their intervention status.

Primary outcome

The primary outcome for the trial is a science attainment test developed by the University of York (Joshi et. al 2021, updated 2022). The test measures KS2 science ability across biology, chemistry and physics and includes questions on "working scientifically".[7] The test is made of 15 questions in total, with 38 parts and a maximum score of 45 marks. The questions include multiple choice, free text, drawing and graphical responses.

The science attainment test used in the Focus4TAPS efficacy trial was refined and developed during the course of the efficacy trial[8]. This version was developed based on quantitative testing and qualitative feedback, including consideration of coverage of key curriculum concepts (Joshi et al., 2021). It demonstrated high internal reliability (Cronbach's α = .87), and factor analysis findings suggested a single latent variable underlying the scores. The test also showed expected patterns with Key Stage 1 data and across known groups of students.[9]

The tests will be administered to pupils in June and July 2025 using a paper booklet. The tests will be invigilated by an external party, AlphaPlus, to ensure that pupils work alone. AlphaPlus are an external organisation who will conduct invigilation and

[7] Focus4TAPS is centred on Working Scientifically in the National Curriculum which, in Year 5, involves planning different types of investigations and controlling variables, taking accurate and precise measurements using a range of equipment, recording data using increasingly complex graphs and diagrams, interpreting and reporting results, and linking them to scientific evidence.
[8] Focus for Teacher Assessment of Primary Science (Focus4TAPS) evaluation report, EEF, 2022
[9] The original version of the Thinking, Doing, Talking Science (TDTS) attainment test (Hanley, Slavin, and Elliott, 2015) was designed to cover aspects of the primary National Curriculum as there has been no widely recognised standardised science assessment for English primary pupils since 2008. The original version was found to be a poor match with the national curriculum in a comparison commissioned by the EEF. The adapted version used in the efficacy trial was developed as a refined version that correlates between with the national curriculum.

marking of the science attainment test. Pupils will have 45 minutes to complete the test. Invigilators will not be made aware of the school's randomly allocated group (Focus4TAPS or teaching as usual). The completed tests will be processed by an external company. The tests will then be double marked by two independent markers using the mark scheme provided by the developer. Any discrepancies between the first two marks will be reviewed by a third independent marker. Data processors and markers will all be unaware of the Focus4TAPS or control status of any tests they are handling.

Secondary outcomes

The secondary outcome is pupils' attitudes and beliefs towards science, measured using a questionnaire developed by the evaluation team for the Focus4TAPS efficacy trial. Development of the measure was informed by existing national and international instruments and the measure consists of seven subscales:

(1) Interest and enjoyment in science
(2) Confidence in science
(3) Perceptions of science teachers
(4) Self-regulation of learning in science
(5) Self-efficacy for Working Scientifically
(6) Working Scientifically beliefs
(7) Wider benefits of science.

The questionnaire includes 47 questions across the seven subscales with responses to each question recorded on a four-point Likert scale: 1 ('Disagree a lot'), 2 ('Disagree a little'), 3 ('Agree a little'), and 4 ('Agree a lot'). In the efficacy trial, each subscale had a Cronbach's α value ranging from 0.7 to 0.9. Questionnaires will be provided as paper booklets for pupils to complete. The completed questionnaires will be processed by the NatCen data team, who will be unaware of whether a pupil has received the Focus4TAPS intervention or not. The questionnaire was administered to pupils at baseline and will be administered again in June and July 2025, alongside the science attainment test.

Baseline measures

The baseline measures used for the primary outcome analysis will be Key Stage 1 maths and reading classifications. These classifications will be requested from the NPD and are categorical with the following categories:

- A = Absent
- D = Disapplied
- BLW = Below
- PKF = Pre-key stage standards
- WTS = Working towards the expected standard
- EXS = Working at the expected standard
- GDS = Working at a greater depth within the expected standard

To control for prior KS1 attainment in the analysis, we will create separate binary variables indicating KS1 maths and reading attainment classifications. The category 'EXS' and 'GDS' will be combined and categorised to take the value 1. In alignment with the approach used in the efficacy trial, the categories 'A' and 'D' will be identified as missing, as this reflects prior attainment being unavailable.

These baseline measures were chosen because Key Stage 1 is the latest standardised examination that Year 5 pupils will have taken by the start of the intervention. The measure is standardised across schools and available through the NPD. In the efficacy trial, the correlation coefficient between KS1 measures and science attainment (the primary outcome variable) amounted to 0.529, which is relatively high. These scores will be requested from the National Pupil Database (NPD) to characterise pupils' attainment at age 6 to 7. It is anticipated that baseline scores will explain around 28% of variance in outcomes, as this was the percentage of the variance KS1 measures explained in science attainment in the efficacy trial.

The secondary outcome analysis will include baseline measures for each of the beliefs and attitudes to science subscales. Each subscale will be treated as a separate outcome and analysed individually. This data was collected at baseline (completed between 11 September and 14 October 2024) and will improve power as the baseline and endline scores are expected to be correlated.[10]

Sample size calculations overview

The evaluation protocol anticipated the following sample sizes:

- A total of 260 schools would be recruited into the trial, with half randomly allocated to the Focus4TAPS intervention group and the other half to the control group.
- On average, around 25 pupils were expected in one Year 5 class per school, consistent with sample sizes observed in the efficacy trial.

The recruitment strategy was designed with the expectation that around 15% of schools would likely withdraw from the trial after initial recruitment, and that around 16% of pupils would either withdraw from the trial or miss the endline test and survey by the end of the trial. These expected levels of school- and pupil-level attrition were therefore expected to result in a sample size of 221 schools, averaging around 21 pupils per school. Total planned sample size was expected to be around 4,641 pupils by the end of the trial.

Power calculations were conducted using the PowerUp! Excel tool (Dong and Maynard, 2013). The estimated Minimum Detectable Effect Size (MDES) for the primary outcome, associated with the planned sample size in the protocol, was 0.182 standard deviations. This is the smallest possible change in outcomes that the trial

[10] We do not have estimates of correlation between baseline and endline for similar measures as similar EEF trials (such as Thinking, Doing Talking Science or other primary science trials) as well as other similar existing research tended to use standardised test measures to proxy for the outcome measure at baseline.

would be adequately powered to detect and is within the threshold of 0.2 required for the trial to potentially receive a '5-padlock' rating by the EEF.

Achieved sample sizes

This SAP was written after randomisation was completed. Randomisation was carried out for 240 schools on 03 July 2024 as several schools withdrew or had not uploaded pupil data by the time of randomisation, corresponding to 7.7% attrition at the school level between recruitment and randomisation.

At the time of writing the SAP (31 March 2025), five schools, all allocated to receive the Focus4TAPS intervention, had withdrawn from the trial after randomisation. This brings the number of schools remaining in the trial to 235 (115 Focus4TAPS intervention, 120 control). This represents a 2.1% attrition rate at the school level. Reasons for withdrawal included staffing changes (n = 1), various circumstances (n = 1), and unspecified reasons (n = 3).

Updated sample size calculations

Table 4 presents updated sample size calculations for this trial. These calculations indicate the smallest effect size (MDES), measured in standard deviations, that the trial is likely to be able to detect, given the sample sizes available and the underlying assumptions.

The table presents sample size and MDES estimates across all key stages of the trial: recruitment, randomisation, protocol, SAP writing, and the expected sample at the analysis stage.

The sample size calculations presented in the table reflect different assumptions at each stage of the trial. Pre-attrition MDES estimates are reported for the recruited sample of schools (n = 260) and the sample of schools that were randomised (n=240). Post-attrition MDES are reported at the protocol stage, based on a 15% attrition rate at the school level and 16% at the pupil level, relative to the recruited sample. The table also includes updated power calculations at the time of writing the SAP, taking into account the total number of schools remaining active in the study and the numbers of pupils included in the data uploaded by all schools.[11] Post-attrition MDES are reported for the expected sample at analysis, applying the original attrition assumptions to the pupil data available at the time of SAP writing.

The calculations use the same assumptions as those in the evaluation protocol, using parameters informed by the efficacy trial findings. We make the conservative assumption that school-level stratification variables do not explain any variance in outcomes. Pupil pre-test / post-test correlations from the efficacy trial are used in the calculations. We assume a Type I error rate of 0.05 and a Type II error rate of 0.20 (power of 0.80).The intra-cluster correlations (ICCs) used in the calculations (0.204 overall, 0.248 FSM) are based on the final analysis of the efficacy trial, where

[11] Pupil data was uploaded for 6,371 pupils across 232 schools at the time of writing the SAP (31 March 2025). Three schools had not yet uploaded pupil data, of which two of these had participated in baseline data collection activities. These schools have been included in the number of schools used in the effect size calculations. The harmonic mean used in the calculations are calculated from the uploaded pupil data.

observed clustering effects were higher than initially assumed at protocol and randomisation stage (Mutjaba et al., 2022).

Table 4: Minimum detectable effect sizes for intention-to-treat analysis

Stage		Recruitment		Randomisation		Protocol writing		SAP writing		Expected sample at analysis stage	
Attrition assumptions		*School attrition: 0% Pupil attrition: 0%		*School attrition: 0% Pupil attrition: 0%		*School attrition: 15% Pupil attrition: 16%		*School attrition: 0% Pupil attrition: 0%		*School attrition: 15% Pupil attrition: 16%	
		OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.166	0.204	0.173	0.212	0.182	0.227	0.175	0.220	0.182	0.236
Pre-test/ post-test correlations	level 1 (pupil)	0.529	0.501	0.529	0.501	0.529	0.501	0.529	0.501	0.529	0.501
	level 2 (school)	0	0	0	0	0	0	0	0	0	0
Intracluster correlations (ICCs)	level 2 (school)	0.204	0.248	0.204	0.248	0.204	0.248	0.204	0.248	0.204	0.248
Alpha		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2	2	2	2	2
Average cluster size		25	6	25	6	21	5	25	5	21	4
Number of schools	intervention	130	130	120	120	111	111	115	115	111	111
	control	130	130	120	120	110	110	120	120	110	110
	total	260	260	240	240	221	221	235	235	221	221
Number of pupils	intervention	3250	780	3000	720	2331	555	2875	575	2331	444
	control	3250	780	3000	720	2310	550	3000	600	2310	440
	total	6500	1560	6000	1440	4641	1105	5875	1175	4641	884

***Notes*:** Power calculations were conducted in *PowerUp!* (Dong and Maynard, 2013) using standard assumptions (alpha = 0.05, power = 0.80). The software requires estimates of the proportion of variance explained through the included pupil- and school-level covariates (R^2). These were calculated as the squares of the pre-test/post-test correlations; pupil-level R^2 was taken as 0.28 (0.529^2) for the full sample and 0.25 (0.501^2) for the FSM subgroup. Power calculations for the recruitment, randomisation, and protocol stages were based on efficacy trial estimates. Specifically, average cluster sizes were based on the sample sizes of pupils included per school in the efficacy trial. Power calculations for SAP writing and for the expected sample at analysis stage are based on actual sample sizes in the data uploaded by schools at the time of writing (31 March 2025). At SAP writing, average cluster sizes were calculated as the harmonic mean of the number of pupils per school (25 overall, 5 FSM). For the expected sample at analysis stage, a 16% pupil-level attrition rate was applied to the SAP data, resulting in revised average cluster sizes of 21 pupils overall and 4 pupils for the FSM subgroup. The expected number of schools at analysis stage reflects a 15% school-level attrition rate from the recruited sample of 260 schools, resulting in a final estimated total of 221 schools.

Given sample sizes at randomisation, the trial had 80% power to detect an impact of 0.173 standard deviations in the overall sample, and 0.226 standard deviations in the FSM sample. The MDES of 0.173 is larger than for the full sample of recruited schools (reflecting school-level attrition of 7.7% between recruitment and randomisation), but lower than the post-attrition MDES expected at the protocol writing stage (0.182). This suggests that the trial met the MDES assumptions outlined at protocol, and sample sizes are in line with expectations for the overall sample. Sample sizes for the FSM subgroup were lower than expected at protocol, highlighting that the study may not be well-powered to detect a small impact (of 0.2 standard deviations) for the FSM sample.

At SAP writing, pupil data uploaded by schools was used to calculate the harmonic mean number of pupils per school. This was 25 for the overall sample and 5 for the FSM subgroup. Given these sample sizes, the trial has 80% power to detect an effect size of 0.175 standard deviations in the overall sample and 0.220 standard deviations in the FSM sample. This indicates that the trial remained well powered at the SAP stage, with MDES estimates aligned with expectations set out in the protocol.

For the expected sample at analysis stage, a 16% pupil attrition rate was applied to sample sizes available in the pupil upload data at the SAP writing stage. This resulted in a revised average of 21 pupils per school in the overall sample and 4 pupils in the FSM subgroup. The number of schools was assumed to follow the expected 15% attrition rate for the recruited sample of schools, as assumed at protocol writing. Under these assumptions, the MDES was estimated at 0.182 for the overall sample and 0.236 for the FSM subgroup. This suggests that the trial is still likely to detect an effect size of 0.2 standard deviations in the overall sample but may not have sufficient statistical power to detect small standardised effects of 0.2 standard deviations among the FSM subgroup.

Analysis

Primary outcome analysis

IEQ1. What is the impact of schools' participation in the Focus4TAPS programme on the science attainment of Year 5 pupils (as measured by the science attainment test), compared with teaching as usual?

The primary outcome analysis will use an Intention-to-treat (ITT) approach to estimate the impact of the Focus4TAPS programme on the science attainment of Year 5 pupils. The model will be a two-level linear regression, with pupils at level one and schools at level two, as well as a random intercept by school to allow for school-level random effects and within-school clustering effects.

In addition to baseline measures of KS1 attainment, a binary indicator of treatment allocation, the randomisation strata and pupil age will also be included as covariates in the primary outcome analysis. The model will include pupils age in months to account for the developmental effect reported in the materials for the attainment test as well as to improve efficiency of estimation. This implies the following basic model:

$$ScienceAttainment_{ij} = \beta_0 + \beta_1 KS1Maths_{ij} + \beta_2 KS1Reading_{ij} + \beta_3 Intervention_j + Strata_j'\beta_4' + \beta_5 age_{ij} + u_j + e_{ij} \quad (1)$$

where pupils (i) are nested within schools (j). β_0 is an overall intercept, $KS1Maths_{ij}$ and $KS1Reading_{ij}$ are baseline KS1 maths and reading attainment indicators, and the intervention effect is captured by β_3. The $Strata_j$ variables represent the stratification blocks created by the stratification criteria used during randomisation[12], while age_{ij} is age in months. The stratification variables are included to reflect the stratified randomisation. Age in months is included as a covariate to account for the developmental effect on child performance reported in the development materials for the attainment test. The term u_j is a school-level random effect and e_{ij} is the error term, both assumed to be normally distributed and uncorrelated with all the covariates included in the model.

The impact of the Focus4TAPS programme will be expressed as an effect size using Hedges' *g* formula. See the Effect size calculation section below for an explanation of how effect sizes will be calculated. Following EEF statistical analysis guidance (EEF, 2022), we will also present histograms of the pre- and post-test scores, along with a summary of means and standard deviations of pre- and post-test scores. The analysis will be conducted in R using the `lmer()` function in the `lme4` package.[13] [14]

Secondary outcome analysis

IEQ2. What is the impact of the Focus4TAPS programme on Year 5 pupils' attitudes towards science compared to teaching as usual?

The secondary outcomes will be pupils' attitudes towards science measured by the seven subscales from the attitudes to science questionnaire. An ITT approach will be used with separate multilevel linear regression models (with school level random intercepts and other covariates included in the analysis of the primary outcome) fitted for each subscale score[15] including, as an additional covariate, the corresponding science subscale score at baseline. As outlined in the secondary outcomes section, the attitudes to science questionnaire is administered at both baseline and endline. A baseline measure will be included in the model to improve power as the baseline and endline subscale scores are expected to be correlated.

The basic form of the model for each subscale is as follows:

[12] Schools were stratified by geographic region and school level KS1 maths and reading attainment.

[13] As the baseline KS1 reading and maths attainment will be supplied from NPD, the analysis will need to be conducted through the Office for National Statistics Secure Research Service (ONS SRS). R 4.3 is the most up to date version available in the SRS environment at the time of this SAP being written. Version information is available at Supporting your research project - Office for National Statistics (ons.gov.uk) [Assessed 28/03/2025]

[14] https://doi.org/10.32614/CRAN.package.lme4

[15] Each subscale was rated on a 4-point Likert scale (1 = Disagree a lot to 4 = Agree a lot). The possible score ranges for each subscale are as follows: Interest and enjoyment in science (9 items): 9–36; Confidence in science (7 items): 7–28; Perceptions of science teachers (10 items): 10–40; Self-regulation of learning in science (5 items): 5–20; Self-efficacy for Working Scientifically (6 items): 6–24; Working Scientifically beliefs (6 items): 6–24; Wider benefits of science (4 items): 4–16.

$$SubscaleScore_{ij,k} = \beta_0 + \beta_1 BaselineSubscaleScore_{ij,k} + \beta_2 Intervention_j + Strata_j'\beta_3' + \beta_4 age_{ij} + u_j + e_{ij} \quad (2)$$

where pupils (i) are nested within schools (j), and k specifies the particular subscale. β_0 is an overall intercept, $BaselineSubscaleScore_{ij,k}$ is the corresponding subscale score at baseline, and the intervention effect is estimated by β_2. The $Strata_j$variables represents stratification blocks, while age_{ij} is pupil age in months. The term $u_j$is a school-level random effect and e_{ij} is the error term.

Due to multiple secondary outcomes being analysed, 95% confidence intervals and continuous p-values for these outcomes will be reported before and after Romano-Wolf step-down p-value adjustment, which considers the greater likelihood of Type I errors while allowing for dependence among outcomes in the data (Romano & Wolf, 2005; 2016).

Subgroup analyses

Subgroup analysis will be conducted both for primary and secondary outcomes.

IEQ3. Is any impact (IEQ1 and IEQ2) moderated by eligibility for Free School Meals (FSM)?

For IEQ3, we will assess whether Focus4TAPS has a differential impact for pupils who are eligible for FSM and received the Focus4TAPS programme, compared to pupils who are eligible for FSM and received teaching as usual. In order to carry out this analysis, FSM eligibility will be based on the variable "EVERFSM_6_P" in the NPD, which indicates if a pupil has been recorded as eligible for FSM at any time in the last 6 years.

For the FSM subgroup analysis, two approaches will be undertaken. Firstly, we will use a similar model to that applied for the primary analysis (equation 1) on the restricted subsample of individuals who are eligible for FSM. The β_3 coefficient in this model will therefore provide the estimated treatment effect specifically for the FSM subgroup.

Secondly, we will add an interaction term between the intervention and the binary indicator to indicate FSM eligibility. The following interaction model will be run using all pupils in the analysis:

$$ScienceAttainment_{ij} = \beta_0 + \beta_1 KS1Maths_{ij} + \beta_2 KS1Reading_{ij} + \beta_3 Intervention_j + \beta_4 FSM_{ij} + \beta_5 Intervention_j \cdot FSM_{ij} + Strata_j'\beta_6' + \beta_7 age_{ij} + u_j + e_{ij} \quad (3)$$

where pupils (i) are nested within schools(j). β_0 is an overall intercept, $KS1Maths_{ij}$ and $KS1Reading_{ij}$ are baseline KS1 maths and reading attainment classifications for each pupil, and the intervention effect is captured by β_3. The β_4 coefficient represents the differential attainment for pupils eligible for FSM and the coefficient on the interaction term (β_5) captures the differential impact of Focus4TAPS for the FSM-eligible subgroup. The $Strata_j$ variables represent the stratification blocks, while age_{ij}

is pupil age in months. The term u_j is a school-level random effect and e_{ij} is the error term. The results from the two approaches will be compared as a sensitivity check.

The above analysis will be repeated for each of the secondary outcome subscales separately to assess whether the impact of the Focus4TAPS intervention varied by FSM eligibility. Analogous to the analysis of IEQ2, 95% confidence intervals and continuous p-values for these outcomes will be reported before and after Romano-Wolf step-down p-value adjustment, which consider the greater likelihood of Type I errors and the dependence among these secondary outcomes.

IEQ4. Is any impact (IEQ1 and IEQ2) moderated by any of the following characteristics? (a) gender (b) exposure to science (c) KS1 maths category (d) pupils' pre-intervention interest in science

For IEQ4, we will assess whether Focus4TAPS had a differential impact for pupils looking at the following characteristics:

1. Gender (as indicated in the NPD, or where missing in the NPD, as self-reported by pupils via the survey questionnaire)[16]
2. KS1 maths category (from the NPD)
3. Pupil pre-intervention interest in science (as reported by pupils in the baseline pupil survey questionnaire)
4. Exposure to science (derived from a set of questions asked in the baseline pupil survey questionnaire, namely (i) whether pupils' parents/carers attended university; (ii) whether pupils have a family member who works as a scientist or in a job using science or medicine; (iii) pupils thinking that their parents/carers are interested in science; (iv) pupils having teachers with STEM backgrounds).

Pupil pre-intervention interest in science will be taken from the *Interest and enjoyment in science* subscale in the baseline pupil survey data. The subscale consists of nine items rated on a 4-point Likert scale. A mean score will be calculated for each pupil and the scoring of responses for any negatively orientated items (such as 'Science is boring') will be reversed for consistency within this process. A binary variable will be derived from this score: pupils with scores at or above the median for the sample will be coded as one (high pre-intervention interest) and pupils below the median as zero (low pre- intervention interest).[17]

Exposure to science will be derived from the above list of survey questions by applying a dimension reduction approach, using factor analysis to combine the four variables of interest into a single variable. A binary indicator of 'exposure to science' will then be derived from the factor variable, which will be equal to one if the value of the factor variable is above the median for the sample, and zero otherwise.

[16] Pupil gender will be taken from the NPD where available. Where gender data is missing in the NPD, self-reported gender from the attitudes and beliefs to science questionnaire will be used to maximise completeness and minimise data loss in the analysis.

[17] Baseline data on the secondary pupil survey outcomes had been collected at the timing of writing the SAP. The nine items included in the pupil interest in science measure had high completion rates of 98-99%, with 94% of pupils completing all nine items.

For each of the above subgroups, we will run interaction models with interaction terms between the intervention and the subgroup binary indicators – for both primary and secondary outcomes. This implies the following model for each subgroup (example shown for the primary outcome):

$$ScienceAttainment_{ij} = \beta_0 + \beta_1 KS1Maths_{ij} + \beta_2 KS1Reading_{ij} + \beta_3 Intervention_j + \beta_4 Subgroup_{ij} + \beta_5 Intervention_j \cdot FSM_{ij} + Strata'_j\beta'_6 + \beta_7 age_{ij} + u_j + e_{ij} \quad (4)$$

Similar to the interpretation of equation (3) above, the β_4 coefficient represents the differential attainment for the subgroup of interest and the interaction term (β_5) captures the differential impact of Focus4TAPS for this subgroup. A significant β_5 coefficient on the interaction term is indicative of there being a differential effect across the subgroups considered.

Following the approach in the efficacy trial, where a statistically significant interaction effect is revealed in the interaction model analysis, separate models will be estimated for the subgroups separately. Effect sizes by subgroup will therefore be reported only for subgroup-outcome combinations where the interaction model demonstrates statistically significant effects. Given the high number of estimates generated by the subgroup analysis, multiple-hypothesis testing will be used by applying Romano-Wolf correction within each subgroup, as described in the secondary analysis section.[18]

Longitudinal follow-up analyses

Long term follow-up for this trial is currently not considered.

Imbalance at baseline

We will explore potential imbalance at baseline between the Focus4TAPS programme and teaching as usual group following baseline assessment.

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

- Gender
- FSM eligibility
- Baseline KS1 reading attainment
- Baseline KS1 maths attainment
- Baseline attitudes and beliefs to science subscale scores

Potential imbalance for gender, FSM eligibility and baseline KS1 reading and maths attainment will be checked with cross-tabulations, including a count and percentage by group allocation. A standardised mean difference threshold of 0.1 will be used to assess baseline imbalance. Differences of 0.1 or more will be considered as an indication of possible imbalance.

Baseline attitudes to science subscale scores will be summarised with descriptive statistics (n, mean, standard deviation, range, median and effect sizes) by group

[18] The high number of hypotheses being tested increases the risk of Type I error – with eight outcomes considered per subgroup, the probability of at least one false positive finding within the eight interaction models included in the subgroup increases to 34% (Ranganathan and others, 2016). Where separate restricted sample models are run for each subgroup, three models would be run for each of the eight subscales, indicating at least 24 tests – this increases the probability of at least one false positive to about 70%.

allocation and differences in science subscale scores reported as Hedges' *g* effect sizes.[19] The criteria for possible imbalance for the subscales will be set at an effect size of 0.05 or more. As imbalance on this dimension will be captured only for pupils with baseline survey data, we will include in reporting the degree of baseline survey non-response, and describe how this may affect the assessment of imbalance.

In addition to exploring potential baseline imbalance at the individual-level, the distributions of the following school-level characteristics will also be presented for Focus4TAPS and teaching as usual allocation.

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

- Geographic region
- Proportion of pupils eligible for FSM
- School-level KS1 maths attainment
- School-level KS1 reading attainment

Sensitivity analysis and robustness checks

Where imbalance is indicated, a sensitivity analysis will be estimated to capture the effect of imbalanced covariates on the primary outcome. This will involve re-running the primary outcome analysis model (equation 1) controlling for the covariates that were imbalanced at baseline. We will compare the effect size estimates across both models to understand potential changes.

Additional robustness checks will also include estimating White standard errors using the `vcovCR()` function in the `clubSandwich` package in R, which adjusts for clustering and heteroskedasticity in the data.[20]

We will also report Cronbach's alpha estimates for both the primary and secondary outcomes to indicate whether the science attainment test, and the individual subscales used in the secondary outcomes all perform as expected. We will compare these estimates with those obtained in the efficacy trial to assess test validity.

Missing data

We will follow the protocol for missing data suggested by the EEF (see EEF, 2022). In the case of missing data, we will first analyse the prevalence and pattern of missingness in outcomes and pre-intervention covariates descriptively with cross-tabulations including counts and percentages in each category. If less than 5% of the data is missing, no further action will be taken, and the primary analysis will be run on complete cases only.

If the missing rate exceeds 5%, an assessment of the pattern of missingness will be carried out by regressing the binary missingness indicator of the science attainment

[19] Baseline attitudes and beliefs regarding science are expected to capture dimensions including exposure to science and pupils' interest in science.
[20] https://doi.org/10.32614/CRAN.package.clubSandwich

score on a set of relevant covariates including gender, age, FSM eligibility, KS1 maths and reading attainment, and baseline attitudes and beliefs in science.

This analysis of the pattern of missingness will use a multilevel logistic regression model. If the pattern of missingness is unrelated to the intervention effect (e.g. absence due to pupil illness), then missing data will be assumed missing completely at random (MCAR) and we will continue with a complete case analysis.

A multilevel logistic regression model will be used to assess whether covariates are considered missing at random conditional on any (other) covariates. The outcome for this model will be a binary indicator taking the value one if any covariates are missing, and zero otherwise. If a covariate is considered missing at random conditional on other covariates using this model, the primary analysis will be re-estimated through multiple imputation (MI) using chained equations (MICE) and results will be presented alongside the complete data analysis as part of sensitivity analysis. Multiple imputation will be conducted using the `mice` package in R.[21]

If missing data appears to be missing not at random, multiple imputation will be used as described earlier, in addition to some sensitivity analyses (Carpenter and others, 2007).

Given the large number of secondary outcomes, missing data analysis will only be undertaken on the primary outcome.

Compliance

IEQ5: What is the Complier Average Causal Effect (CACE) of the Focus4TAPS programme on the science attainment of Year 5 pupils (as measured by the science attainment test), compared with teaching as usual?

A Complier Average Causal Effect (CACE) analysis will be undertaken to show the impact of the Focus4TAPS programme on the primary outcome (science attainment) compared to individuals receiving teaching as usual, taking into account level of compliance with Focus4TAPS.

Compliance in this trial is defined at the school-level and two criterions will capture compliance to Focus4TAPS:

1. Attendance of train-the-trainer sessions by the Year 5 teacher
2. Delivery of assessment class lessons

As shown in Table 5, for a school to be compliant the Year 5 teacher needs to attend at least 2 out of 3 sessions of training and deliver a minimum of five assessment class lessons using Focus4TAPS by the end of the academic year.

Table 5: Compliance measure indicators

Number	Criteria	Data source	Indicator
1	Attendance of train-the trainer sessions	Attendance register	Binary indicator taking the value 1 if the Year 5

[21] https://doi.org/10.32614/CRAN.package.mice

			teacher attends at least 2 out of 3 sessions of training; 0 otherwise.
2	Delivery of lessons	Teacher self-report in endline survey	Binary indicator taking the value 1 if at least 5 assessment class lessons using Focus4TAPS delivered by the end of the year[22]; 0 otherwise.

**Notes:* Missing data on any criterion will be scored as zero (that is, the setting will be considered non-compliant on that criteria).

The compliance measure will be defined as a categorical measure taking values 0, 1 or 2 corresponding to the counts of compliance criteria that the school fulfils. This measure has a maximum value of 2, and is calculated as the sum of compliance criteria indicators for each school. This therefore allows for partial compliance to indicate how impact estimates vary by level of compliance among Focus4TAPS intervention schools.

Where there is imperfect compliance across schools, a CACE analysis will be undertaken, drawing on an instrumental variable (IV) to estimate the CACE for schools where the intervention allocation was complied with, with standard errors clustered at the school level. Specifically, the CACE will estimate the effect of the Focus4TAPS intervention on science attainment outcomes for pupils in schools that complied with the intervention allocation.

In order for compliance analysis to be conducted using an instrumental variables approach, instrument allocation has to be a good 'instrument' for compliance. This depends on the following conditions being met:

- Intervention allocation being highly correlated with compliance. This can be assessed using the F-statistic from the first stage results in the two-stage least squares approach, which will be reported.
- Intervention allocation only affects science attainment through compliance and does not directly affect science attainment. This would be expected as schools being randomly allocated to the Focus4TAPS intervention group would determine whether teachers are eligible to attend training.

Initial group allocation will be used as an instrumental variable for the compliance indicator in a two-stage least squares (2SLS) estimation approach (Imbens and Angrist, 1994), therefore adjusting the impact estimates by the compliance rate.

The first stage of analysis will involve the compliance indicator (as a continuous dependent variable) regressed on all covariates that are used in the main primary

[22] The intervention protocol stipulates a requirement for teachers to deliver at least five Focused Assessment lessons, but there is no maximum number of lessons that teachers can deliver. There are more than 40 lessons available for teachers to choose from.

outcome model and will include, as an IV, a binary variable that indicates schools' pre-intervention treatment allocation. The first stage equation is as follows:

$$Compliance_j = \alpha_0 + \alpha_1 KS1Maths_{ij} + \alpha_2 KS1Reading_{ij} + \alpha_3 Intervention_{ij} + Strata_j'\alpha_4' + \alpha_5 age_{ij} + u_j + v_{ij} \quad (5)$$

The first-stage results will be reported alongside the correlation between the intervention allocation and the compliance indicator (α_3) and the associated F-statistic to indicate instrument relevance.

The second stage of analysis will regress science attainment on the predicted compliance estimated in stage 1. The second stage equation is as follows:

$$ScienceAttainment_{ij} = \beta_0 + \beta_1 KS1Maths_{ij} + \beta_2 KS1Reading_{ij} + Strata_j'\beta_3' + \beta_4 age_{ij} + \beta_5 \widehat{Compliance}_j + u_j + e_{ij} \quad (6)$$

The coefficient on the second-stage predicted compliance measure (β_5) is the CACE estimate of the compliance effect. In the event that there are no confounding factors affecting compliance and attainment, the CACE estimate will be equal to the intention-to-treat estimate.

The compliance analysis will be estimated only for the primary outcome of science attainment.

Intra-cluster correlations (ICCs)

The intra-cluster correlation (ICC) is a measure of how strongly pupils from the same school resemble one another in terms of their achieved scores. If the ICC is high, this indicates that pupils from the same school are likely to score similarly on the science assessment and attitudes to science questionnaire.

The ICCs will be estimated directly from the primary and secondary outcomes analysis models, using the variance estimates for each level of clustering. The ICC for schools $\boldsymbol{\rho_S}$ will be estimated using the following formula based on Hedges (2011):

$$\boldsymbol{\rho_s} = \frac{\boldsymbol{\sigma_{BS}^2}}{\boldsymbol{\sigma_{BS}^2} + \boldsymbol{\sigma_{WS}^2}} = \frac{\boldsymbol{\sigma_{BS}^2}}{\boldsymbol{\sigma_{WT}^2}}$$

where $\boldsymbol{\sigma_{BS}^2}$ is the between-school variance, $\boldsymbol{\sigma_{WS}^2}$ is the within-school variance and $\boldsymbol{\sigma_{WT}^2}$ the total variance.

Effect size calculation

We will calculate the effect sizes (ES) for cluster-randomised trials as adapted from Hedges (2007)[23]:

$$ES = \frac{(\overline{Y_T} - \overline{Y_C})_{adjusted}}{\sqrt{\sigma_s^2 + \sigma_e^2}}$$

Where $(\overline{Y_T} - \overline{Y_C})_{adjusted}$ is the mean difference between the Focus4TAPS and teaching as usual group adjusted for baseline characteristics, while $\sqrt{\sigma_s^2 + \sigma_e^2}$ is an estimate of the population standard deviation. σ_s^2 is the variance of school level intercept and σ_e^2 is variance of residuals.

We will take each group's adjusted mean and variance to calculate the effect size. The variance used will be the pooled variance (across both pupil and school setting, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors).

A 95% CI for the ES, that takes into account the clustering of pupils in schools, will also be reported. The CI will be based on the standard error of the adjusted effect size, following the approach outlined by Hedges (2007) for cluster-randomised trials. Specifically, the CI will be calculated as:

$$ES \pm 1.96 \times SE(ES)$$

where *SE(ES)* accounts for the variance due to clustering and is derived from the multilevel model residual variance and the number of clusters.

For the FSM subgroup analysis, ES will be calculated using two approaches. First, we will calculate the ES using the restricted sample of pupils eligible for FSM, following the same approach as the overall ES calculation above. Secondly, as a sensitivity check, we will calculate ES for FSM pupils using an interaction model, which captures the adjusted intervention effect for FSM pupils:

$$ES_{FSM} = \frac{\beta_1 Intervention_j + \beta_5 (Intervention_j * FSMEver_{ij})}{sd}$$

where $\beta_1$represents the main treatment effect for non-FSM pupils, and β_5 captures the differential effect for FSM pupils. The standard deviation (sd) will incorporate both within-school (σ_e^2) and between-school (σ_s^2) variance to adjust for clustering effects. The ES from the interaction effects model will be compared to the ES derived from the restricted sample and reported in the appendix.

Effect sizes will be calculated for each of the models estimated.

[23] Hedges, L. V. (2007) 'Effect Sizes in Cluster-Randomized Designs' Journal of Educational and Behavioural Statistics 32(4): 341–370

References

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259–275.

Dong, N., & Maynard, R. (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), 24–67.

Education Endowment Foundation [EEF]. (2022). *Statistical analysis guidance for EEF evaluations*.

Hanley, P., Slavin, R., & Elliott, L. (2015). *Thinking, Doing, Talking Science: Evaluation report and Executive summary*. London: Education Endowment Foundation.

Hedges, L. V. (2007). Effect Sizes in Cluster-Randomized Designs. *Journal of Educational and Behavioural Statistics,* 32(4): 341–370

Imbens, G. W., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica*, *62*(2), 467–475.

Joshi, K. et al., 2021. *Development of a Year 5 Science Assessment for the EEF*, London: Education Endowment Foundation

Leonard, H., Poquillon, G., Bhatti, S., Morgan, H., Benny, L., & Wadsworth, K. (2024). *Focus4TAPS Evaluation Protocol*. https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/focus4taps_-_evaluation_protocol.pdf?v=1720421081

Mutjaba et al. (2022). Focus for Teacher Assessment of Primary Science (Focus4TAPS) Evaluation report. Available at Focus4TAPS-Report.pdf (d2tic4wvo1iusb.cloudfront.net)

Appendix A: Randomisation syntax

The code below shows the Stata syntax used to conduct the randomisation of 240 schools into the Focus4TAPS intervention and control groups. Randomisation was stratified by region and KS1 attainment from the 2021/22 academic year, when the current Year 5 pupils completed their KS1 assessments. A fixed random seed, based on the date of randomisation (03724), was used to ensure replicability. The researcher carrying out the randomisation was blinded to any school-identifying information at the time of randomisation.

```
version 17.0

*Importing school data

use "F4t School data blinded.dta"

*Sorting by strata variables: region and KS1 attainment

sort region ks1_attainment

*Randomly assigning schools to two groups within each stratum.
*The date of randomisation will be used as seed.

randtreat, generate(treatment) replace strata(region ks1_attainment)
misfits(global) setseed(03724)

*Labelling groups

lab def treat 0 "Control" 1 "Treatment", replace

lab values treatment treat
```

Appendix B: Randomisation allocation within strata

Strata	Control	Treatment	Total
Region			
Bury	9	9	18
Below median	6	7	13
Above median	1	1	2
Missing KS1 scores	2	1	3
Derby	8	8	16
Below median	5	5	10
Above median	2	3	5
Missing KS1 scores	1	-	1
Essex	8	8	16
Below median	2	2	4
Above median	5	6	11
Missing KS1 scores	1		1
Liverpool	7	8	15
Below median	3	4	7
Above median	4	3	7
Missing KS1 scores	-	1	1
Newcastle	17	16	33
Below median	6	6	12
Above median	8	8	16
Missing KS1 scores	3	2	5
Nottingham	5	6	11
Below median	3	4	7
Above median	2	2	4
Missing KS1 scores	-	-	-
SE London	9	8	17
Below median	2	2	4
Above median	6	6	12
Missing KS1 scores	1	-	1
Sheffield	11	12	23
Below median	5	6	11
Above median	6	5	11
Missing KS1 scores	-	1	1
Southampton	7	7	14
Below median	4	4	8
Above median	3	2	5
Missing KS1 scores	-	1	1
Suffolk	9	9	18
Below median	4	3	7
Above median	4	4	8

Missing KS1 scores	1	2	3
Sunderland	12	12	24
Below median	6	5	11
Above median	5	5	10
Missing KS1 scores	1	2	3
W London	7	7	14
Below median	2	2	4
Above median	5	5	10
Missing KS1 scores	7	7	14
Wolverhampton	11	10	21
Below median	4	4	8
Above median	4	4	8
Missing KS1 scores	3	2	5
All regions	**120**	**120**	**240**
Below median	52	54	106
Above median	55	54	109
Missing KS1 scores	13	12	25
Total	**120**	**120**	**240**