Thinking, Doing, Talking Science (TDTS) Statistical Analysis Plan

Evaluator (institution): University of York

Principal investigators: Lyn Robinson-Smith/David Torgerson (previously Pam Hanley/Louise Elliott)

PROJECT TITLE	Thinking, Doing, Talking Science (second re-grant - a two-armed, cluster randomised trial)
DEVELOPER (INSTITUTION)	Science Oxford (The Oxford Trust)
EVALUATOR (INSTITUTION)	York Trials Unit, University of York
PRINCIPAL INVESTIGATORS	Louise Elliott (up to July 2022), Dr Pam Hanley (up to December 2022) From December 2022: Dr Lyn Robinson-Smith, Prof David Torgerson
PROTOCOL AUTHORS	Pam Hanley, Louise Elliott, Imogen Fountain, Jenny Roche, Laura Mandefield, Caroline Fairhurst, Dr Lyn Robinson-Smith
SAP AUTHOR	Caroline Fairhurst
TRIAL DESIGN	Two-arm, cluster randomised controlled trial with random allocation at school level: *Cohort 1:* Year 5 (main trial) - 2022-23 Year 6 (longitudinal follow-up) - 2023-24 *Cohort 2:* Year 5 - 2023-24
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	Key Stage 2 9-10 years; Year 5 (main trial) 10-11 years; Year 6 (longitudinal follow-up)
NUMBER OF SCHOOLS	Planned: 180 primary schools[1] Actual: 180 schools
NUMBER OF PUPILS	Planned: 8,100 per cohort, i.e. 16,200 over the 2 years

[1] or middle schools if they include both Year 5 and Year 6

	Actual: (estimated) Cohort 1 = 168 schools, 7,239 pupils
PRIMARY OUTCOME MEASURE AND SOURCE	Science attainment at the end of Year 5 (Cohort 1 only) Year 5 Science Assessment, 15-item measure scored 0-45, Centre for Industry Education Collaboration (CIEC) and York Trials Unit (YTU), University of York
SECONDARY OUTCOME MEASURE AND SOURCE	1. Science Attitudes Science Attitudes Questionnaire, 27-item measure, 5-point Likert scale, based on Kind, Jones & Barmby, 2007, standard score from total score 20-100 (Cohort 1 and 2) 2. Science attainment at Year 6 Year 6 Science Assessment, in development (Cohort 1) 3. Science attainment at Year 5 Year 5 Science Assessment, 15-item measure scored 0-45, CIEC and YTU (Cohort 2) 4. Key Stage 2 (Reading (KS2_READSCORE, range 0-120), Maths (KS2_MATSCORE, range 0-120)) (Cohort 1)

This analysis plan was written post-randomisation but prior to receipt of any outcome data and deals only with the statistical analysis of effectiveness for the main trial. This document has been written based on information in the study protocol version 1.1 dated 04/04/2023, published on the EEF website in which full details of the background and design of the trial are presented.

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0	04/04/2023	*N/A. Creation of original document.*

Table of contents

Introduction

The primary science experience heavily influences subsequent subject attitudes but is often low priority and teachers may lack confidence teaching it (Harlen & Qualter, 2008; Slavin et al, 2014). Thinking, Doing, Talking Science (TDTS) is a continuing professional development (CPD) programme for teachers that aims to enable the teachers to adapt their pedagogy to plan and teach creative science lessons that overtly encourage their pupils' higher order thinking.

Further details of the trial rationale and the TDTS intervention can be found in the trial protocol.

Design overview

This two-armed cluster randomised controlled trial (RCT) with random allocation at school level will evaluate the effect of TDTS on science attainment at the end of Year 5.

This trial runs across two years. The first year forms the main trial. Year 5 teachers will attend CPD sessions in the academic year 2022-23, four of which will be spread throughout the first two terms, with a further half-day in the third (Summer) term. Teachers will be given 'gap' tasks/strategies to use with their classes between the sessions and encouraged to reflect on their implementation, discuss with their in-school colleagues and then feedback at the next CPD session. Any Year 5 teachers who join the school during the year should inherit the previous teacher's file and receive input from the other participating teacher(s) in their school as well as attending any subsequent training sessions, to reflect the real-world approach.

The second year of the trial (with a second cohort of Year 5 pupils that will be recruited in the 2023-24 academic year) will examine the 'legacy' of the TDTS training and any effects of embedding of the TDTS practices. Schools will be encouraged, wherever possible (e.g. unless the teacher is no longer at the school or operational circumstances make it impossible) to ensure that the same teachers will be retained in Year 5 for both years of the evaluation. No training will be provided by the TDTS team to teachers new to Year 5 in the second trial year, but the final half-day of TDTS training will include a section on cascading the approach to colleagues across the school. The intervention schools in this second year of the evaluation will therefore have a mix of teachers that taught a Year 5 class at an intervention school in the first year and/or received training from the TDTS team, and teachers new to TDTS who did not teach a Year 5 class in an intervention school in the first year and have received no external training in TDTS (but may have received cascade training from an experienced teacher at their school).

The second year will also follow the first cohort of Year 5 pupils into Year 6 to assess the 'legacy' effects of exposure to the TDTS programme. Year 6 teachers may have received TDTS training if they have moved from a Year 5 class the previous year or it has been cascaded within the school but no training will be provided by TDTS to Year 6 teachers.

Research questions for the first cohort are:

Main trial: Cohort 1 – Year 5 / Year 6

RQ 1. What is the impact of the TDTS programme, in comparison to usual Year 5 provision, on the science attainment of Year 5 pupils? *[primary outcome]*

RQ 2. What is the impact of the TDTS programme, in comparison to usual Year 5 provision, on pupils' attitudes towards science? *[secondary outcome]*

RQ 3. What is the impact of the TDTS programme, in comparison to usual Year 5 provision, on the science attainment of Year 5 pupils who are eligible for Free School Meals (FSM)? *[subgroup analysis]*

RQ 4. What is the long-term impact of the TDTS programme, in comparison to usual Year 5 provision, on pupils' science attainment at the end of Year 6 and on Key Stage 2 outcomes (Year 6 SATs attainment in Reading and Maths)? *[secondary outcomes]*

Research questions for the second cohort are:

Second year: Cohort 2 – Year 5

RQ 5. What is the impact of the TDTS programme, in comparison to usual Year 5 provision, on the science attainment of Year 5 pupils given the mix of experienced and inexperienced teachers in the intervention group? *[secondary outcome]*

RQ 6. What is the impact of the TDTS programme on pupils' attitudes towards science, in comparison to usual Year 5 provision, given the mix of experienced and inexperienced teachers in the intervention group? *[secondary outcome]*

RQ 7. What is the impact of the TDTS programme, in comparison to usual Year 5 provision, on the science attainment of Year 5 pupils who are eligible for FSM given the mix of experienced and inexperienced teachers in the intervention group? *[subgroup analysis]*

Research questions 1-3 will be answered by analyses due to be conducted in Autumn 2023 and written up in a report to be submitted to the EEF in Spring 2024.

Research questions 4-7 will be answered by analyses due to be conducted in Autumn 2024 and written up in an addendum report to be submitted to the EEF in Spring 2025.

Table 1: Trial design overview

Trial design, including number of arms	Two-arm, cluster randomised, 2 cohorts. Cohort 1 followed for 2 years: Year 5 2022-23 to Year 6 2023-24 Cohort 2 followed for 1 year: Year 5 2023-24
Unit of randomisation	School
Minimisation variables (if applicable)	Geographical region (6 levels: Lancashire, Lincolnshire and East Midlands, North East, South West, Staffordshire and West Midlands, Yorkshire) Percentage of pupils eligible for free school meals in the school (taken at the time of recruitment from the latest census data) (2 levels: dichotomised at the median <24%; ≥24%)

Primary outcome	variable	Science attainment at the end of Year 5 (Cohort 1 only)
	measure (instrument, scale, source)	Year 5 Science Assessment, 15-item measure scored 0-45, Centre for Industry Education Collaboration (CIEC) and York Trials Unit (YTU), University of York
Secondary outcome(s)	variable(s)	Attitudes towards Science Science attainment Attainment in Mathematics and Reading
	measure(s) (instrument, scale, source)	**Cohort 1:** *At the end of Year 5:* Science Attitudes Questionnaire, 27-item measure, 5-point Likert scale, based on Kind, Jones & Barmby, 2007 (standard score from total score 20-100) *At the end of Year 6:* Year 6 Science Assessment, YTU (currently under development, scoring to be confirmed) Key Stage 2 (Year 6 SATs attainment in Reading and Maths) from the National Pupil Database: • English Reading (KS2_READSCORE, range 0-120) • Maths (KS2_MATSCORE, range 0-120) **Cohort 2:** *At the end of Year 5:* Year 5 Science Assessment, 15-item measure scored 0-45, CIEC and YTU (standard score from total score 20-100) Science Attitudes Questionnaire, 27-item measure, 5-point Likert scale, based on Kind, Jones & Barmby, 2007
Measurement of Prior Attainment	variable	Early Years Foundation Stage Profile (EYFSP)
	measure (instrument, scale, source)	Average EYFSP point score obtained by combining all 17 Early Learning Goals (ELG), scored 1-3, NPD

Outcome measures

Baseline measure

To minimise costs and the burden on schools it was decided to use existing data available in the National Pupil Database (NPD). The baseline measure for all analyses will be the average point score from the 17 Early Learning Goals (ELGs) that make up the Early Years Foundation Stage Profile (EYFSP). This baseline measure has been chosen as an alternative to the Key Stage (KS) 1 English (Reading) and Mathematics scores used in the previous effectiveness trial as KS1 results are not available for the cohorts of pupils in this trial, who would have been in Year 2 during the academic year 2019-20 (Cohort 1) or 2020-21 (Cohort 2) when national KS1 assessments were cancelled due to the COVID-19 pandemic. KS1 results would have been the preferred choice for the baseline measure as this would have allowed a direct comparison of results with the previous effectiveness trial, and it is likely that the correlation between KS1 results and the outcomes in the trial would have been higher than with EYFSP results as these were assessed longer ago.

Within the EYFSP (for the academic years of 2019-18 and 2018-19 when the pupils in this trial would have been in Reception), for each ELG, the child's learning and development was rated as:

- Best described by the level of development expected at the end of the EYFS (expected)
- Not yet at the level of development expected at the end of the EYFS (emerging)
- Beyond the level of development expected at the end of the EYFS (exceeding)

These will be scored as scored 1 = emerging, 2 = expected, 3 = exceeding, and all 17 scores will be summed and averaged (to produce a total score ranging from 1-3).

The EYFSP will be obtained in Autumn 2023 for cohort 1 and Autumn 2024 for cohort 2.

Primary outcome

Science Attainment

The measure used for both the efficacy (Hanley et al, 2015) and effectiveness (Kitmitto et al, 2018) trials of TDTS is no longer fit for purpose. Its creation (Abrahams et al, 2014) preceded the new science curriculum (DfE, 2013) with its changed content and emphases (e.g. more focus on "working scientifically"/science enquiry). The main alternative (GL Progress Test in Science) is not considered to be a varied enough test (for instance, it is predominantly multiple choice) to be an adequate replacement. Therefore, we will use a new measure, the Year 5 Science Assessment, recently developed by the Centre for Industry Education Collaboration (CIEC) and York Trials Unit (YTU), University of York (Joshi et al, 2022) and designed to be suitable to be administered to Year 5 pupils as a meaningful outcome measure in future evaluations. It was originally developed for use in two EEF-funded RCTs in 2020. However, both these trials were delayed because of school closures due to the Covid-19 pandemic; therefore, it has not yet been used in any published trial. This new measure has been developed to better reflect the current curriculum, have a mix of question types and have greater emphasis on "working scientifically" than the alternatives. Details of the development and validation of this measure are published in Joshi et al, 2022.

This is a 15-item measure, each item is worth between 1 and 5 marks (three items are worth 1 mark, one item is worth 2 marks, seven items are worth 3 marks, one item is worth 4 marks, and three items are worth 5 marks), and incomplete items are given a score of 0. Item scores are summed to produce a total score from 0 to 45.

The primary outcome analysis will be based on the Cohort 1 Year 5 results. Invigilators, recruited and trained by the Evaluation Team, will administer the tests within schools. Invigilators will be blinded as far as possible (e.g. there could be instances where teachers mention TDTS if they are in the intervention group, but this shall be discouraged).

Secondary outcomes

Science attitudes (Cohort 1 and 2)

The science attitudes instrument used in both the efficacy trial (Hanley et al, 2015) and the previous effectiveness trial (Kitmitto et al, 2018) contained 23 items asking about interest, self-efficacy and activity in science lessons. In the TDTS pre-trial four new items were added to the instrument to strengthen the self-efficacy scale. This 27-item, self-reported science attitudes questionnaire was administered to 103 pupils from two schools in the pre-trial. Each item is scored from 5 = agree a lot to 1 = disagree a lot (with negatively worded items reverse scored). Factor analysis on data from the pre-trial indicated that 20 of these items can be incorporated into a scale that measures ‘interest and self-efficacy’ (to be published in TDTS pre-trial report). The 27-item scale will be completed in-class supervised by class teachers, at the end of Year 5 for both cohorts in the TDTS main trial. Responses to the 20 items identified by the factor analysis will be summed to generate a total score from 20-100, where a higher score indicates a greater interest in science. The score will be standardised to a mean of 0 and a standard deviation (SD) of 1 by subtracting the sample mean from each pupil’s score and dividing it by the sample standard deviation. The remaining items, not used in this scale will be summarised separately.

Pragmatics dictate the attitudes survey will be teacher-administered, rather than being completed with the trained invigilators during visits to complete the primary outcome, because otherwise the session would be too long for pupils of this age - 45+ minutes for the science assessment plus this survey. Teachers will be given instructions about how to administer the science attitudes questionnaire (they will facilitate a session where the students complete the survey). This is the way it has been done in the two previous trials.

Science Attainment (Cohort 2)

The same 15-item science attainment test used for the primary outcome will be administered to Cohort 2 at the end of Year 5.

Longitudinal outcomes for Cohort 1

English Reading and Maths

We will assess for any impact on Maths and English on Cohort 1 at the end of Year 6 by considering attainment based on pupils’ KS2 results (English Reading and Maths), which will be obtained from the NPD in Autumn 2024. These will be measured via scaled assessment scores, using the variables KS2_READSCORE and KS2_MATSCORE, both scored on a scale from 0-120.

Science Assessment

At the end of Year 6 for Cohort 1, we intend to collect the secondary outcome of science attainment, assessed via a new measure, the Year 6 Science Assessment, currently being developed by the YTU. This new measure will reflect the current curriculum, have a mix of question types and an emphasis on “working scientifically”. Further details will be added to this SAP when the report for the development for this measure is published.

Randomisation

A statistician at YTU randomised schools 1:1 to either the intervention arm (offered the TDTS CPD programme) or the control arm (continue with usual provision for the duration of the evaluation).

A dedicated computer program, MinimPy (Saghaei and Saghaei, 2011), was used for randomisation via minimisation using the following factors:

- **School region – 6 levels**: Lancashire, Lincolnshire and East Midlands, North East, South West, Staffordshire and West Midlands, and Yorkshire for logistical reasons, to ensure a balanced spread of intervention and control schools in each area.
- **School deprivation level** (i.e. the percentage of pupils eligible for free school meals in the school based on latest available data) – 2 levels: dichotomised at the median for the 171 schools that were randomised in the first batch (see below) <24%, ≥24%) to ensure balance between the randomised groups, since this school characteristic and individual child deprivation may moderate outcomes.

Randomisation was carried out in batches (groups of schools that were ready to be randomised at that time) to avoid delays in programme induction and to maximise programme delivery for as many schools as possible. Deterministic (i.e. without the use of a random element) minimisation was used (Altman and Bland, 2005). This was deemed to be sufficient as the allocations were conducted in batches, rather than one-by-one prospectively, meaning predictability was not a concern and hence a random element was not required.

In total, 180 settings were randomised (90 Intervention, 90 control) in two ‘batches’, one of size 171 and a second of size nine. Schools were randomised and informed of their allocation at the end of the academic year 2021-22 in order that intervention settings could arrange to send their teachers to the training sessions; however, pupil details and baseline data collection could only be completed at the start of the academic year 2022-23 when schools knew which pupils would be in the class.

Sample size calculations overview

Table 2: Sample size calculations

		Protocol		Randomisation	
		OVERALL	FSM	OVERALL	FSM
Minimum Detectable Effect Size (MDES)		0.15	0.19	0.16	0.20
Pre-test/ post-test correlations	level 1 (pupil)	0.5	0.5	0.5	0.5
	level 2 (class)	-	-	-	-
	level 3 (school)	-	-	-	-
Intracluster correlations (ICCs)	level 2 (class)	-	-	-	-
	level 3 (school)	0.15	0.15	0.15	0.15
Alpha[2]		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		45	~8	43.1	7.5
Number of schools	intervention	90	90	89	89
	control	90	90	79	79
	total	180	180	168*	168*
Number of pupils	intervention	4,050	700	3,790	656
	control	4,050	700	3,449	597
	total	8,100	1,400	7,239	1,253

*180 schools were randomised, but 12 withdrew before providing pupil details

From protocol

A summary of the assumptions used in the calculation of the sample size are given in Table 2, and full details provided in the trial protocol. The trial is powered to detect an effect size of 0.15 in the primary outcome (science attainment for Cohort 1 at the end of Year 5) with 80% power and two-sided alpha of 0.05, assuming pupil-level attrition of 15%, a pre- and post-test correlation of 0.5, an intracluster correlation coefficient (ICC) of 0.15 and an average year group (cluster) size of 45 at randomisation. A total of 180 schools are required (8,100 pupils per year group).

[2] Accounting for 15% attrition

As of January 2020, 17.3% of pupils were eligible for FSM. Assuming we recruit 180 schools and an anticipated total of 8,100 pupils per each year of the trial, there will be approximately 1,400 pupils eligible for FSM each year (approximately 8 per school). Under the same assumptions as above, an MDES of 0.19 will be detectable. All calculations were conducted in Stata (Version 15, StataCorp. 2017).

At randomisation

At randomisation, there were 180 settings. As mentioned above, for logistical reasons, schools had to be randomised and informed of their allocation at the end of the academic year 2021-22 so intervention schools could begin to make arrangements to attend the training. However, we could only collect pupil details from participating schools at the start of the academic year 2022-23. Twelve schools withdrew during this process, and so the number of schools from which we received participating pupil details was 168 (7,239 pupils[3]). This is an average of 43.1 pupils per school. Assuming 80% power, two-sided 5% alpha, a pre- and post-test correlation of 0.5, an ICC of 0.15, and 15% pupil-level attrition, the MDES with this sample size would be approximately 0.16.

Approximately 1,253 of the randomised pupils will be eligible for FSM (average of ~7.5 per school). With this sample size, under the same assumptions, the MDES would be approximately 0.20 for this subgroup.

[3] This figure is subject to change as data are finalised.

Analysis[4]

Analysis will follow the EEF's (2022) most recent guidance. The trial statistician will not be blind to group allocation. All analysis will be conducted in Stata version 17 (StataCorp. 2021), or later (to be confirmed in final report). All analyses will be conducted using the principles of intention to treat (ITT) including all schools and pupils in the groups that they were randomised to, irrespective of whether or not they went on to receive the intervention.

Statistical significance will be assessed using two-sided tests at the 5% significance level. Estimates of effect with 95% confidence intervals (CIs) and p-values will be provided.

A CONSORT diagram will be produced to show the flow of schools and children through the trial. The number of children identified as eligible for the evaluation and the numbers actually assessed at baseline and post-test will be reported with reasons for non-participation given where available.

All outcome data will be summarised descriptively by trial arm. Effect sizes based on the adjusted difference between the groups at the outcome assessment point will be presented as unadjusted and adjusted mean differences with their associated 95% confidence interval and p-value. Treatment effects will also be presented as Hedges' g effect sizes, and converted to estimated additional months' progress (Higgins et al. 2015).

The EEF analysis guidance states that the ITT population should exclude any pupils and school that dropped-out after randomisation, but before allocation is revealed; we do not have any such cases in this trial.

The correlation between average EYFSP score and all outcomes (separately) will be presented, as will the correlation between science attitude and attainment outcomes for each year. Histograms of pre and post-test scores will be presented.

Imbalance at baseline

School and pupil characteristics and outcome measures at baseline will be summarised descriptively by randomised group, both as randomised (to check the randomisation achieved balance) and as analysed in the primary analysis (to check whether attrition has introduced selection bias into the complete-case sample). At school level, the following data will be summarised: geographical region (Lancashire, Lincolnshire and East Midlands, North East, South West, Staffordshire and West Midlands, Yorkshire), percentage of pupils eligible for FSM, whether the school is urban or rural, type of school, and latest Ofsted rating. At child level, the following data will be summarised: gender, FSM status (from NPD variable EVERFSM_6_P) and EYFSP measure of prior attainment.

Continuous measures will be reported as a mean, standard deviation (SD) (and/or median, minimum and maximum) while categorical data will be reported as a count and percentage. No hypothesis testing of comparisons of the baseline data between the two groups will be undertaken (de Boer et al., 2015). The difference in prior attainment (EYFSP score) between the groups will be reported as the Hedges' g effect size, with a 95% CI.

Primary outcome measure and analysis

The primary analysis will investigate any difference in science test scores (Year 5 Science Assessment) between the two arms. A linear mixed effects regression model at the pupil-level

[4] Please see the Statistical Analysis Guidance.

will be used to estimate the adjusted mean difference in scores. School will be included as a random effect and group allocation, average EYFSP score and the minimisation factors (region, FSM) will be included as fixed effects. FSM is dichotomised at the school level for use as a minimisation factor in the randomisation, but will be entered into the analysis model as a dichotomous variable at the pupil level (EVERFSM_6_P from the NPD) as this provides more granular information. Robust standard errors will be specified to account for any potential heteroscedasticity.

Model equation:

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 I_{Ai} + \beta_3 I_{Bi} + \beta_4 I_{Ci} + \beta_5 I_{Di} + \beta_6 I_{Ei} + \beta_7 FSM_{ij} + \beta_8 I_{Gi} + u_i + y_{ij}$$

Y_{ij} = response (science attainment) of the j-th of n_i members of the i-th cluster (school), *$i=1,\ldots,m$, $j=1,\ldots,n_i$*
m = number of clusters (school)
n_i = size of cluster (school) i
x_{ij} = baseline EYFSP score for j-th member of i-th cluster (school)
I_{Ai} = indicator variable for location of i-th school (1 = Lancashire, 0 = elsewhere) (Yorkshire is reference category)
I_{Bi} = indicator variable for location of i-th school (1 = Lincolnshire and East Midlands, 0=elsewhere)
I_{Ci} = indicator variable for location of i-th school (1 = North East, 0 = elsewhere)
I_{Di} = indicator variable for location of i-th school (1 = South West, 0 = elsewhere)
I_{Ei} = indicator variable for location of i-th school (1 = Staffordshire and West Midlands, 0 = elsewhere)
FSM_{ij} = indicator variable for EVERFSM_6_P status for j-th member of i-th cluster (school) (no FSM is reference category)
I_{Gi} = indicator variable for group allocation of i-th cluster (school) (0 = Control, 1 = Intervention)
β_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7, β_8 = fixed effect parameters
$u_i \sim N(0, \varphi_b^2)$ = setting-specific random effect and $\gamma_{ii} \sim N(0, \varphi_w^2)$ = individual-specific random effect

The coefficient for the group allocation indicator (β_8) will represent the effect estimate of TDTS with respect to the primary outcome.

The normality of the standardised residuals will be checked using a QQ plot. If the model assumptions are in doubt, a sensitivity analysis will be conducted in which a log transformation of the outcome and/or covariate data will be tried to improve the model fit. If needed, other transformations (e.g. square root) will be tried.

Subgroup analyses

A subgroup analysis will be conducted for the primary outcome of Year 5 science attainment considering FSM status (EVERFSM_6_P), first by retaining the whole analytic sample and including an interaction between FSM and group allocation in the primary analysis model, and secondly by repeating the primary analysis only within the restricted FSM subgroup. This will be conducted for both cohorts separately.

Missing data

The amount of missing baseline and outcome data will be summarised, and reasons for missing data will be explored and provided in the report where available. Where less than 5%

of ITT pupils are missing from the primary analysis model, no further action will be taken. If the percentage of missing cases exceeds 5%, then multi-level logistic regression models will be used to model presence or absence of the primary outcome (binary variable indicating missingness) adjusting for all available pupil and school-level baseline data as fixed effects, and school as a random effect. Significant predictors and possible mechanisms for the missing data will be discussed in the report. The presence of baseline factors that are significantly associated with missingness provides evidence that the data are missing at random.

Additionally, the impact of missing data on the primary analysis (if >5%) will be assessed using multilevel imputation, which assumes the data are missing at random. This will be done via the REACOM-Impute macro, which is compatible with Stata (http://www.bristol.ac.uk/cmm/software/realcom/imputation.html), by including all available pupil and school-level baseline variables (school: location, percentage of pupils eligible for FSM, whether the school is urban or rural, type of school, latest Ofsted rating; pupil: gender, FSM status, EYFSP measure of prior attainment). This imputation procedure can account for the two-level (pupil and school) nature of the data.

A ‘burn-in’ of 10 will be used, which means that the first 10 iterations of the imputation are not used to allow the iterations to converge to a stationary distribution, and 30 imputed datasets will be created (the values of 10 and 30 are subject to the convergence of the model and other values may be used during analysis). The primary analyses will then be re-run within the imputed datasets and Rubin’s rules (Rubin, 1987) will be used to combine the multiply imputed estimates. This analysis will be repeated for both cohorts separately.

Compliance

Compliance will be measured as a binary outcome at class level rather than school level. The Delivery Team keep registers of attendance at each training session, which they provide to the Evaluation Team. When schools send their participating pupils details, they indicate what class the child is in and who their teacher is. Then teacher changes throughout the year are recorded by the Delivery Team and passed to the Evaluation Team.

Definitions of compliance:

Cohort 1 - first year of trial: *The class has been taught by a teacher who attended at least 3 out of the 4 full days of training.*

This would include a class that (because of long-term sick leave, resignations etc.) has been taught by two teachers who together have attended 3+ training days. For example, a class would be considered compliant if the Year 5 teacher attends two days training in the Autumn term then leaves the school; then the new teacher attends at least one further full-day TDTS training session.

Cohort 2 - second year of trial: *The predominant teacher of the class attended at least 3 full days of training in the first year of the trial.*

The predominant teacher will be defined as the teacher who taught the class for the majority of the academic year based on termly updates from each school.

A CACE analysis (Dunn, Maracy and Tomenson, 2005) for each cohort will be conducted for the Year 5 science attainment outcome using the dichotomous compliance measures described above. These analyses will use a Two Stage Least Square (2SLS) approach with group allocation as the instrumental variable for the compliance indicator, with cluster standard errors to account for clustering at the school level. Results for the first stage (of the 2SLS process) are reported alongside i) the correlation between the instrument and the endogenous variable (presented as the partial R^2 statistic from the first-stage estimation); and, ii) a F test (F statistic and p-value). The F statistic should exceed 10 for inference based on the 2SLS estimator to be reliable when there is one endogenous regressor, as in this case (Bound et al., 1995; Stock et al., 2002).

Secondary outcome measures and analysis

The secondary outcome of Year 5 science attainment for Cohort 2 will be analysed as described for the primary outcome.

Scores from the 'interest and self-efficacy' scale of the science attitudes questionnaire (at the end of Year 5) will be compared between the two trial arms. As for the primary analysis, a linear mixed effects regression model will be used to estimate the adjusted mean difference in scores. School will be included as a random effect and group allocation, average EYFSP score and minimisation factors (as in the primary analysis) will be included as fixed effects. This will be conducted in both cohorts separately.

Longitudinal outcome analysis

The secondary outcomes of Science, Maths and Reading attainment assessed at the end of Year 6 will be analysed similarly to the primary outcome. However, whereas for science the objective would be to consider whether the intervention had a positive effect, for Key Stage 2 Maths and English outcomes the objective is to check for 'no negative impact', as it would not be necessarily expected that the intervention would have a positive impact. The rationale for this is based on results of analysis of longer-term outcomes carried out by researchers at Durham University at the end of the first effectiveness trial (Singh et al. 2019). They investigated the impact on KS2 Maths and English (average score for two subjects) among the pupils at the end of Year 6 and found a non-significant, but negative effect (-0.07, 95% CI -0.23 to 0.09). Therefore, we propose to interpret results of the analysis of Maths and English in a non-inferiority, rather than a superiority, framework. This means that focus will be on the magnitude of the lower limit of the 95% confidence interval for the treatment effect size, rather than the p-value. We want to ensure that the intervention does not have an overly negative impact on these outcomes.

Additional analysis

As part of the protocol for the development of the Year 6 Science attainment test (Hanley et al., 2022), there is the proposal to investigate the correlation between test scores from the Year 5 and Year 6 tests. It was decided that this should be conducted as part of the TDTS trial. Total Science attainment test scores from Year 5 pupils in Summer 2023 will be correlated with the total test scores from Year 6 pupils in Summer 2024, presented using Pearson's Correlation Coefficient and a 95% confidence interval (using the control arm only).

Intracluster correlation coefficients (ICCs)

The ICC associated with school for the outcomes (both pre and post-test where available) will be presented alongside a 95% CI. The ICC at post-test will be computed for the analysis model, and also for an empty model (i.e., one without covariates). The ICC at pre-test will be calculated for a linear model with pre-test as the outcome and setting as a random effect.

Effect size calculation

Hedges' g effect sizes will be calculated by dividing the adjusted mean difference between the intervention and control group (accounting for baseline measures and the minimisation factors) by the pooled unconditional standard deviation obtained from the model run without these covariates. A 95% CI for the effect size will be calculated by dividing the 95% confidence limits for the adjusted mean difference by this same denominator. All parameters used in these calculations will be provided in the final report.

$$\mathrm{ES} = \frac{(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}}{sd_{pooled}}$$

where, $(\overline{\mathrm{Y}}_{\mathrm{T}} - \overline{\mathrm{Y}}_{\mathrm{C}})_{\mathrm{adjusted}}$ denotes the difference in means between trial groups adjusting for pre-test score and the minimisation factors, from the multilevel analysis model; and sd_{pooled} denotes the pooled, unconditional standard deviation of the two groups (square root of the sum of the within- and between-cluster variances) (Xiao et al, 2016).

References

Abrahams, I., Bennett, J., Cheung, A., Elliott, L., Hanley, P., Oberio, Z., ... Turkenburg, M. (2014). Evaluation of the impact of a Continuing Professional Development (CPD) course for Primary Science Specialists: Final report. London: The Wellcome Trust.

Altman DG, Bland JM. Treatment allocation by minimisation. BMJ. 2005 Apr 9;330(7495):843. doi: 10.1136/bmj.330.7495.843. PMID: 15817555; PMCID: PMC556084.

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak. *Journal of the American Statistical Association, 90(430), 443–450.* https://doi.org/10.2307/2291055

de Boer, M.R., Waterlander, W.E., Kuijper, L.D. *et al.* Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate. *Int J Behav Nutr Phys Act* **12**, 4 (2015). https://doi.org/10.1186/s12966-015-0162-z

DfE (2013). The National Curriculum in England: Key Stages 1 and 2 framework document. Retrieved from https://www.gov.uk/government/publications/national-curriculum-in-england-science-programmes-of-study/national-curriculum-in-england-science-programmes-of-study

Dunn, G., Maracy, M. and Tomenson, B. (2005) 'Estimating treatment effects from randomized clinical trials with noncompliance and loss to follow-up: the role of instrumental variable methods', *Statistical Methods in Medical Research*, 14(4), pp. 369–395. doi: 10.1191/0962280205sm403oa.

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

Hanley, P., Elliott, L., Fountain, I., Baird, J., Keding, A., Crompton, Z. (2022). Development of a Year 6 Science Assessment for use as an evaluation outcome measure (protocol). London: Education Endowment Foundation. Retrieved from https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/outcome-measures-and-databases/Y6-science-measure-development-protocol.pdf?v=1670404132

Harlen, W., & Qualter, A. (2008). The teaching of science in primary schools. London: Fulton.

Higgins, S., Katsipataki, M., Coleman, R., Henderson, P., Major, L., Coe, R., & Mason, D. (2015). The Sutton Trust- Education endowment foundation teaching and learning toolkit. Education Endowment Foundation.

Joshi, K., et al. (2022). Development of a Year 5 Science Assessment for the EEF. London: Education Endowment Foundation. Retrieved from https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-outcome-measures-and-databases/year-5-science-assessment

Kitmitto, S., González, R., Mezzanote, J., & Chen, Y. (2018). Thinking, Doing, Talking Science: Evaluation report and executive summary. Available at https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/TDTS.pdf

Kind, P., Jones, K., & Barmby, P. (2007). Developing attitudes towards science measures. International Journal of Science Education, 29(7), 871-893.

Roy, P., Rutt, S., Easton, C., Sims, D., Bradshaw, S. & McNamara, S. (2019). Stop and Think: Learning Counterintuitive Concepts: Evaluation Report. Available at: Stop_and_Think.pdf (educationendowmentfoundation.org.uk)

Rubin DB. Multiple Imputation for Nonresponse in Surveys. Wiley: New York, 1987.

Saghaei, M., & Saghaei, S. (2011). Implementation of an open-source customizable minimization program for allocation of patients to parallel groups in clinical trials. Journal of Biomedical Science and Engineering, 4(11), 734. 35

Slavin, R., Lake, C., Hanley, P., & Thurston, A. (2014). Experimental evaluations of elementary science programs: a best-evidence synthesis. Journal of Research in Science Teaching, 51(7), 870–901.

StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC.

StataCorp. 2021. *Stata Statistical Software: Release 17*. College Station, TX: StataCorp LLC.

Stock, J. H., J. H. Wright, and M. Yogo. 2002. A survey of weak instruments and weak identification in generalized method of moments. Journal of Business and Economic Statistics 20: 518–529.

Xiao Z., Kasim, A., Higgins, S.E. (2016) Same Difference? Understanding Variation in the Estimation of Effect Sizes from Educational Trials International Journal of Educational Research 77: 1-14 http://dx.doi.org/10.1016/j.ijer.2016.02.001