Stop and Think: Learning Counterintuitive Concepts
Statistical Analysis Plan

Evaluator: NatCen Social Research
Principal investigator: Mary McKaskill

Education Endowment Foundation

PROJECT TITLE	**Stop and Think: Learning Counterintuitive Concepts**
DEVELOPER (INSTITUTION)	Birkbeck, University of London; Behavioural Insights Team (BIT)
EVALUATOR (INSTITUTION)	National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR	Mary McKaskill
PROTOCOL AUTHORS	Maha Basharat, Isabel Taylor, Enes Duysak, Tien-Li Kuo
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school-level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	7 – 10, KS2
NUMBER OF SCHOOLS	173
NUMBER OF PUPILS	14,718
PRIMARY OUTCOME MEASURE AND SOURCE	**Outcome:** Maths attainment amongst FSM pupils **Measure and source:** Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment
SECONDARY OUTCOME MEASURE AND SOURCE	**Outcomes:** Maths attainment amongst all pupils Science attainment amongst FSM pupils and amongst all pupils Common misconceptions in Maths and Science **Measure and source:** Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment Year 3: Progress Test in Science (PTM8), 0-40, GL Assessment Year 5: Progress Test in Science (PTM10), 0-50, GL Assessment Both years: Age-specific common misconceptions in Maths and Science tests (Developed by Oxford MeasurEd and NatCen)

SAP version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	05/10/2023	*N/A*

Table of Contents

Abbreviation

2SLS	Two-stage least square
BIT	Behavioural Insights Team
CACE	Complier Average Causal Effect
EEF	Education Endowment Foundation
EYFSP	Early Years Foundation Stage Profile
FSM	Free School Meals
GL PTM	GL Progress Test in Maths
GL PTS	GL Progress Test in Science
ICCs	Intracluster Correlation Coefficient
IPD	Individual Participant Data
ITT	Intention to Treat
IV	Instrumental Variable
KS	Key Stage
MDES	Minimum Detectable Effect Size
MoU	Memorandum of Understanding
NatCen	National Centre for Social Research
NFER	National Foundation for Educational Research
NPD	National Pupil Database
RQ	Research Questions
SAP	Statistical Analysis Plan

Introduction

This statistical analysis plan sets out the intended impact evaluation for the effectiveness trial of Stop and Think. Stop and Think is a computer-assisted programme that aims to improve pupils' ability to adapt to counterintuitive concepts. It does this by training pupils to inhibit their initial, intuitive response and give a slower, more reflective answer instead – in other words, to 'stop and think' about maths and science problems before answering. The programme contents include a series of sessions, made up of questions and multiple-choice answers which include distractors demonstrating common misconceptions. The session topics are aligned to the maths and science curriculum in Years 3 and 5.

The intervention will be delivered by Year 3 and Year 5 teachers in participating primary schools in England. The recipients of the intervention will be the Year 3 and Year 5 pupils, at the start of maths and science lessons (January 2023 and last until May 2023[1]). The intervention is designed to be a whole-class activity, with children working through the problems together as a group. During the delivery period, schools will be expected to deliver a total of 30 Stop and Think sessions, three times per week over a ten-week period. Each session lasts around 12 minutes[2].

The control group in each school will receive teaching as usual. Schools will not be offered financial incentives to participate, as each school will be offered the intervention. However, all schools will receive pupil-level test results via GL Assessment's results portal, as a non-financial incentive to participate in the trial.

The delivery team at Behavioural Insights Team (BIT) recruited primary schools to the intervention and will support teachers to deliver the intervention. This is a change from the efficacy trial[3], when the intervention was delivered by the developer team at Birkbeck.

The evaluation follows a two-arm cluster randomised controlled effectiveness trial of the effect of Stop and Think on Year 3 and Year 5 maths and science attainment. The primary outcome of interest is maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds (as defined as those who have been eligible for Free School Meals (FSM) at any point in the previous 6 years), which is different from that in the efficacy trial. Analysis in the efficacy trial found a positive but non-significant effect of the Stop and Think programme on FSM pupil's Maths attainment in separate models for Year 3 and Year 5 pupils. Furthermore, although this was not included in the published analysis of the efficacy trial, routine post-hoc analysis carried out by the Durham University found a significant, and comparatively large, impact of the intervention for Year 3 and Year 5 FSM-eligible pupils.[4] In addition, addressing pupil disadvantage is a key priority area for EEF. Therefore, maths amongst FSM-eligible pupils was selected as the single primary outcome

[1] This is a change from the efficacy trial, when classroom delivery took place slightly earlier in the academic year (November to March). From a curriculum perspective, this means that pupils will have encountered more of the content covered in the programme by the time delivery starts.
[2] The dosage is unchanged from the efficacy trial.
[3] The efficacy trial was conducted by the National Foundation for Educational Research (NFER) between 2015 and 2018. It had an in-school design with randomisation at the year-group level and was conducted in 89 schools in England. Roy, P. et al. (2019) Stop and Think: Learning counterintuitive concepts. Evaluation report. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop_and_Think.pdf?v=1680163623
[4] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts (137). Unpublished.

The secondary outcomes include maths attainment for all Year 3 and Year 5 pupils, science attainment for all Year 3 and Year 5 pupils, science attainment for Year 3 and Year 5 pupils from disadvantaged backgrounds and the prevalence of common misconceptions in maths and science among all Year 3 and Year 5 pupils.

Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, identical baseline measures for both Year 3 and Year 5 pupils involved in the trial will not be available. The trial will therefore use KS1 maths scores as a measure of prior attainment for pupils in Year 3 and the Early Years Foundation Stage Profile (EYFSP) point score as a measure of prior attainment for pupils in Year 5. More details on these baseline measures are provided in the Stop & Think protocol[5].

Design overview

The impact evaluation is designed as a two-arm cluster randomised controlled effectiveness trial of the effect of Stop and Think on Year 3 and Year 5 maths and science attainment. The full description of the trial is outlined in the protocol[6]. Table 1 summarises the trial design.

Table 1 Trial design

Trial design, including number of arms		Two-arm, cluster randomised control trial
Unit of randomisation		School level
Stratification variables (if applicable)		Class-form entry (whether there is 1, 2 or 3+ classes per year group per year) and the school-level proportion of pupils eligible for FSM (by tercile of distribution)
Primary outcome	variable	Maths attainment amongst FSM pupils
	measure (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment; Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment
Secondary outcome(s)	variable(s)	Maths attainment amongst all pupils Science attainment amongst FSM pupils and amongst all pupils Common misconceptions in maths and science amongst FSM pupils and amongst all pupils
	measure(s) (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment; Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment Year 3: Progress Test in Science (PTS8), 0-40, GL Assessment; Year 5: Progress Test in Science (PTS10), 0-60, GL Assessment Both years: Age-specific common misconceptions in maths and science tests (Developed by Oxford MeasurEd and NatCen)

[5] https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1648826229
[6] https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1648826229

Baseline for primary outcome	**variable**	Year 3: Maths attainment Year 5: EYFSP overall progress[7]
	measure (instrument, scale, source)	Year 3: KS1 maths attainment, 8-category variable ranging from BLW (below expected standard) to GDS (working at a greater depth), National Pupil Database Year 5: Overall EYFSP Point Score, 1-3, National Pupil Database
Baseline for secondary outcome	**variable**	Year 3: Maths attainment Year 5: EYFSP overall progress
	measure (instrument, scale, source)	Year 3: KS1 maths attainment, 8-category variable ranging from BLW (below expected standard) to GDS (working at a greater depth), National Pupil Database Year 5: Overall EYFSP Point Score, 1-3, National Pupil Database

Research questions

This impact evaluation aims to answer the following research questions (RQ):

PRIMARY RESEARCH QUESTION

RQ1. What is the impact of Stop and Think on maths attainment of Year 3 and Year 5 pupils from disadvantaged backgrounds, as measured by FSM status?

SECONDARY RESEARCH QUESTIONS

RQ2. What is the impact of Stop and Think on maths attainment of all Year 3 and Year 5 pupils?

RQ3. What is the impact of Stop and Think on science attainment of all Year 3 and Year 5 pupils?

RQ4. What is the impact of Stop and Think on science attainment of Year 3 and Year 5 pupils from disadvantaged backgrounds, as measured by FSM status?

RQ5. What is the impact of Stop and Think on all Year 3 and Year 5 pupils' misconceptions in maths?

RQ6. What is the impact of Stop and Think on all Year 3 and Year 5 pupils' misconceptions in science?

Randomisation

14,305 schools were invited to take part in the trial, of which 645 schools initially expressed their interest and 410 schools further completed initial phone call. 190 schools returned their Memorandum of Understanding (MoU), of which 17 schools dropped out or did not upload pupil data on time before randomisation. This resulted in a total of 173 schools being included for randomisation to (n = 14,718 pupils).[8]

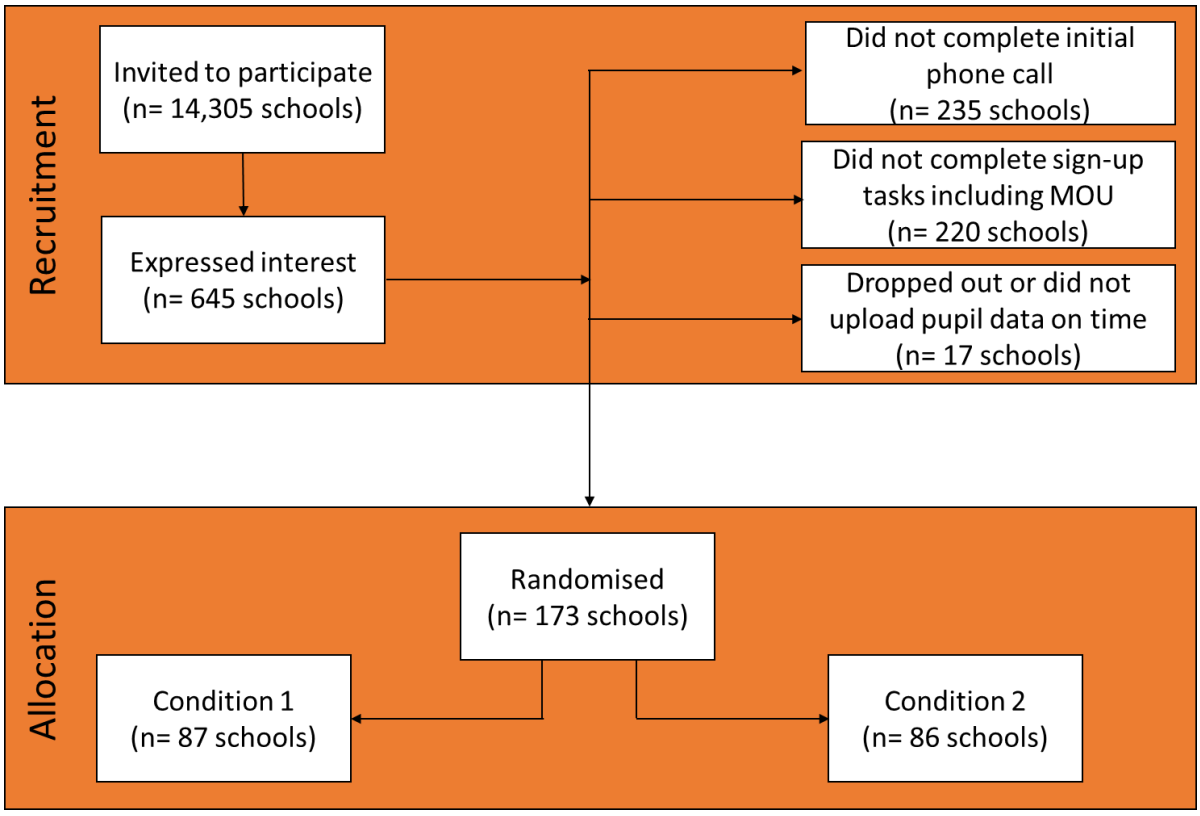
[7] Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, the trial will use the Early Years Foundation Stage Profile (EYFSP) point score as an alternative measure of prior math attainment for pupils in Year 5. Please see the Protocol for details.

[8] After randomisation was conducted, six pupils have withdrawn from the trial by the time of this SAP being written.

Randomisation of settings was carried out blind by the Impact Evaluation team at NatCen using ***randtreat*** command in Stata version 17 on 13 October 2022. Schools involved in the trial were randomly allocated to condition 1 (Year 3 allocated to treatment and Year 5 allocated to control) or condition 2 (Year 5 allocated to treatment and Year 3 allocated to control), such that every school had one intervention and one control year group. The outcome of randomisation was communicated to the delivery team, who in turn notified schools. After the endline data collection, it is understood that the delivery team notified two schools with the opposite condition allocation, though the number of schools by condition remained the same (87 schools in Condition 1 and 86 schools in Condition 2). This is a two-sided non-compliance, which will be dealt with the compliance analysis (see compliance section below for more information).

Figure 1 presents the CONSORT diagram outlining the flow of participating schools from the recruitment to the allocation of conditions. The diagram will be updated in the final report to reflect the flow of schools and pupils from recruitment through randomisation, post intervention assessment and analysis.

Figure 1 Consort diagram



Randomisation at school level was stratified by class-form entry size (whether there is 1, 2 or 3+ classes per year group per year) and the school-level proportion of FSM-eligible pupils (by tercile of distribution). The process is as follows:

- A random number seed was decided and stored;
- Schools were listed in descending order by URN;

- The ***randtreat*** command in Stata was used to randomise schools within each stratum and address misfits globally[9];
- A coin toss was then used to determine which group was assigned to the intervention – if the coin toss is heads, Group A is intervention (and Group B control), and the other way round if the coin toss is tails.

After randomisation, 87 schools were assigned to condition 1 (i.e. Year 3 allocated to treatment and Year 5 allocated to control) and 86 schools to condition 2 (i.e. Year 5 allocated to treatment and Year 3 allocated to control). Table 2 shows the actual allocation of condition within each stratum and Table 3 provides a breakdown of actual allocation of condition by class-form entry and FSM separately.

Table 2 Randomisation allocation across strata (class-form entry size, school-level FSM)

Stratum	Condition 1 (Y3 treat, Y5 control)	Condition 2 (Y5 treat, Y3 control)
1 class, low FSM	14	14
1 class, medium FSM	12	12
1 class, high FSM	13	13
2 class, low FSM	10	10
2 class, medium FSM	12	12
2 class, high FSM	10	10
3+ class, low FSM	6	5
3+ class, medium FSM	4	5
3+ class, high FSM	6	5
Total	87	86

Table 3 Randomisation allocation across class-form entry size and school-level FSM

Stratum	Condition 1 (Y3 treat, Y5 control)	Condition 2 (Y5 treat, Y3 control)
Class-form entry size		
1 class	39	39
2 classes	32	32
3+ classes	16	15
School-level proportion of FSM-eligible pupils		
Low FSM-eligible pupils	30	29
Medium FSM-eligible pupils	28	29
High FSM-eligible pupils	29	28

[9] There were 173 schools to be randomised at this stage. Misfits would occur when observations cannot be evenly distributed among treatment/control groups. We used the global method to deal with misfits. See Carril, A. (2017). Dealing with misfits in random treatment assignment. *The Stata Journal*, *17*(3), 652-667

Randomisation was also conducted at further individual pupil level for outcome measures, for which individual pupils were randomly assigned to sit either maths or science tests. This means that 50% of pupils in each year group will be tested in maths attainment and maths misconceptions and 50% in science attainment and science misconceptions. More details on randomisation are provided in the Stop and Think protocol[10].

Randomisation of test allocation, stratified by school, year group and class, was carried out by the Impact Evaluation team at NatCen in Stata version 17 on 8 March 2023. The procedure followed the same approach as done for school-level randomisation except that schools were listed descending order by unique school identifiers created by NatCen and year group within each stratum. There has been three schools and six pupils dropping out from the trial since settings were randomly allocated into conditions on 13 October 2022. Randomisation of pupil sitting math/science test was therefore carried out with 170 schools and 14,645 pupils. We understand that this level of randomisation (n=14,645 pupils) was conditional on non-attrition as we did not use the full set of pupils as at the point of condition allocation (n=14,718 pupils). This is because we a) would like to ensure balance at the sample across endline testing; b) considered risk of informative attrition as marginal given less than 1% attrition. Table 4 presents actual allocation by subject within each year group.

Table 4 Randomisation of test allocation across year group

Year group	Maths N (% of year group)	Science N (% of year group)
Year 3	3,631 (50%)	3,630 (50%)
Year 5	3,692 (50%)	3,692 (50%)

Primary outcome

The primary outcome for the trial is maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds (defined as pupils eligible for FSM) (RQ1). The primary outcome measure will be an age-standardised measure of pupils' mathematical skills and knowledge - GL Progress Test in Maths (GL PTM)[11] - for pupils eligible for FSM.

MATHS ATTAINMENT (RQ1)

- GL PTM: GL PTM will be used to evaluate Maths attainment for all year 3 and year 5 pupils from disadvantaged backgrounds (defined as pupils eligible for FSM)

Age-appropriate versions of the paper-based test will be delivered to the two year groups (PTM8 for Year 3 and PTM10 for Year 5). Pupils will be assessed in May-July 2023. These tests will be used as no relevant national tests are available for Year 3 and Year 5 pupils through the National Pupil Database (NPD), so GL Progress Test are appropriate age-specific tests for Maths and Science outcomes.

Secondary outcome

The secondary outcomes include maths attainment for all Year 3 and Year 5 pupils (RQ2), science attainment for all Year 3 and Year 5 pupils (RQ3), science attainment for Year 3 and

[10] https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1648826229

[11] For more information, please see https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/.

Year 5 pupils eligible for FSM (RQ4) and the prevalence of common misconceptions in maths (RQ5) and science (RQ6) among all Year 3 and Year 5 pupils.

Maths and science attainment (RQ2-4)

Secondary outcome measures will include GL Progress Test in Math (GL PTM) and GL Progress Test in Science (GL PTS).

- GL PTM: GL PTM will be used to evaluate Maths attainment for all Year 3 and Year 5 pupils.
- GL PTS: GL PTS, will be used to measure science attainment for all Year 3 and Year 5 pupils, and only pupils eligible for FSM.

Age-appropriate versions of these tests will be delivered to the two-year groups (PTS8 for Year 3 and PTS10 for Year 5). Pupils' science and maths attainment will be assessed at the same time, in May-July 2023. As mentioned above, the GL Progress Tests will be used because no relevant national tests are available for Year 3 and Year 5 pupils through the NPD, so GL Progress Test are appropriate age-specific tests for Maths and Science outcomes.

Misconception in maths and science (RQ5-6)

We developed new tests for common misconceptions in maths and science with Oxford MeasurEd to be used as outcome measures in additional models to estimate the effect of Stop and Think on intermediate outcomes. Four different tests were developed for measuring common misconceptions in maths and science (one test per subject per year group). Please see the protocol (Appendix 4) for more details on test development.

Briefly, tests were structured around common KS2 maths and science misconceptions, identified through a literature review. Based on our review, we identified five key misconceptions across curriculum domains for each subject. We also included domains covered by GL Maths and Science Progress Tests, which are being used to measure the primary and secondary outcomes as mentioned above.

We developed and piloted 30 items per test (six pilot items per misconception) to finalise the items for the testsfor each year group and subject (three items per misconception). We expected that pupils would need approximately 45 minutes to complete each 30-item test. The development of misconception test involved a three-stage approach to piloting: a qualitative pre-pilot in five schools, followed by two rounds of validation in 15 schools per round (see Table 5).

Table 5 School and pupil sample by pilot fieldwork stage

	Schools	Year 3 pupils	Year 5 pupils
Qualitative pre-pilot	5	25	25
Validation round 1	15	345	322
Validation round 2	13	273	285
Total pupils		**618**	**607**

Table 6 summarises the number of items, range of scores and reliability summary (Cronbach's α) for validation Round 1 and Round 2. Details on the validity and reliability analysis across validation, including Item Characteristic Curves (ICCs), will be presented in a technical report, published as a standalone output.

Table 6 Summary of misconception test validation by year and subject

	Round 1		Round 2	
	Number of items [range of score]	Master's Partial credit analysis (Cronbach's α)	Number of items [range of score]	Master's Partial credit analysis (Cronbach's α)
Year 3 maths	18 [0,18]	0.7686	14[12] [0,14]	0.6812
Year 5 maths	22 [0,22]	0.7241	15 [0,15]	0.7722
Year 3 science	17 [0,17]	0.6198	16 [0,16]	0.4183
Year 5 science	24 [0,24]	0.5794	16 [0,16]	0.3579

Distractor analysis was carried out to test that wrong answers which contained a common misconception were performing as expected.

Based on the findings from rounds 1 and 2 of analysis, we decided to retain 16 items per test for the Year 3 Maths and Science and Year 5 Maths final tests while we will have only 15 items for Year 5 Science test. In the final report of the Stop and Think effectiveness trial we will use results drawn from the main trial to report:

- Descriptive analysis for misconception scores. These include, the min and max scores, the mean and standard deviation, and the proportion of pupils scored zero and the highest possible scores.
- Cronbach's α and McDonald's ω as part of reliability assessment of the misconception test. These coefficients aim to estimate how well an observed test score measures a construct, given that measurement error produces biased estimates of the associations among constructs that observed variables represent (Bland & Altman, 1997; Flora, 2020).

Data from the misconceptions test will also be used as an intermediate outcome and through the mediation analysis. The results from these analyses will also be presented in the final report. Overall, during endline testing, pupils will take age-appropriate tests (i.e. Year 3 or Year 5 appropriate tests) for common misconceptions in maths or science, depending on their randomised allocation to either maths or science attainment tests as outlined above. The 50% of pupils randomised to take the GL Progress Test in Maths will also take the common misconceptions test in maths. The 50% randomised to take the GL Progress Test in Science will take the common misconceptions test in science. We will use raw misconception scores as misconception outcome measures, whereby the measure is the number of times the learner fell into a common misconception.. The outcome measures of misconception will thus be a count measure ranging between 0-16 for the Year 3 Maths and Science and Year 5 Maths tests and 0-15 for Year 5 Science test. The higher the score is, the higher level of misconception a learner has. The analysis is covered in the section of secondary outcome analyses.

Timeline

[12] Note that one of the fifteen items tested in Round 2 validation had to be removed due to a typographical error in the printed tests. This item has been included in the final version with the error corrected.

Table 7 summarises key dates for the impact evaluation. A full timeline for the effectiveness trial, covering all evaluation activities, can be found in the evaluation protocol[13]. The timeline will also be updated in the final report.

Table 7 Timeline of key dates for the impact evaluation

Date	Activity
December 2021 – July 2022	Recruitment of settings
September – October 2022	Pupil data provided by settings
13 October 2022	Randomisation of settings
13 October 2022	Randomisation information shared with Behavioural Insight Team (BIT), which supports the delivery of the intervention
February– May 2023	Intervention running in schools
March 2023	Randomisation of pupils to maths or science tests
May – July 2023	Endline testing – Year 3 and Year 5
October – December 2023	Impact evaluation analysis
March 2024	Submission of draft EEF report
July 2024	Final EEF report

Sample size calculations overview

This trial is powered to detect a Minimum Detectable Effect Size (MDES) of 0.14 standard deviations for the primary analysis for maths attainment among KS2 pupils eligible for FSM). Details on power calculation are covered in the section of Updated sample size calculations. We have used PowerUp! (Dong and Maynard, 2013) to perform all of the sample size calculations.

Planned sample sizes

The evaluation protocol anticipated the following sample sizes:

- We had aimed to recruit 165 schools in the trial after accounting the school level attrition, with half randomly allocated to Condition 1 and half to Condition 2.
- We assumed an average of 37.2 pupils per year group per school would be recruited to this trial.[14] With an expected pupil attrition of 16%, we also assumed an average of 31.3 pupils per year group per school would be included in the final sample of analysis.

[13] Stop and Think: Learning Counterintuitive Concepts Evaluation Protocol. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1671120133

[14] Based on the pupil retention rate from the efficacy trial of the 'Stop and Think' programme. Roy, P. et al. (2019) Stop and Think: Learning counterintuitive concepts. Evaluation report, p 33. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop_and_Think.pdf?v=1680163623.

- 50% of pupils (an average of 18.6 recruited pupils per year group per school and an average of 15.6 pupils per year group per school for analysis) would be tested each in maths and in science after the intervention is completed.
- Based on publicly available information at the time of protocol writing, we assumed 22.1% of pupils in KS2 in state primary schools in England were eligible for FSM. [15] We therefore estimated an average of 8.2 FSM eligible pupils would be recruited to the trial per year group per school, with an average of 4.1 FSM eligible pupils (as recruited) each tested in maths and in science per year group per school. Given the pupil attrition rate explained above, we estimated an average of 6.9 FSM eligible pupils per year group per school at analysis thus an average of 3.5 FSM eligible pupils per year group per school to be tested in each maths and science.

Achieved sample sizes so far

To allow for school losses after recruitment, we then aimed to recruit 181 schools, assuming that 9% will be lost from the trial after recruitment. Randomisation of schools to conditions was carried out on 13 October 2022, with 173 allocated to condition 1 or condition 2 (see the Randomisation section for details). However, there has been some school-level and pupil-level attrition since then (less than 1%),[16] with 170 schools and 14,645 pupils retained in the sample by the time of pupils being randomly assigned to sit maths/science tests on 8 March 2023. We anticipate that there is likely to be some further attrition by the time of the endline assessments in 2023.

Updated sample size calculations

Tables 8 and 9 present our sample size calculations by maths and science for this trial from the protocol and at randomisation stage of condition allocation. These calculations indicate the smallest effect size, measured in standard deviations, that the trial is able to detect with 80% probability given its sample sizes and a set of underlying assumptions. We will update sample size calculations for the test allocation in the final report.

[15] The assumption was based on previously published figures drawn from the ONS website. The updated FSM figure is available at https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics

[16] Between randomisation of school to conditions and pupils to tests, six individual pupils and three schools (n=67) withdrew from the trial, leaving 170 schools and 14,645 pupils retained in the sample. The pupil-level attrition rate is less than 1%, based on 14,718 pupils at the time of recruitment.

Table 8 Minimum detectable effect size for maths

		Protocol		Randomisation	
		All pupils	FSM (primary analysis)	All pupils	FSM (primary analysis)
Minimum Detectable Effect Size (MDES)		0.13	0.17	0.12	0.14
Pre-test/ post-test correlations	level 1 (pupil)	0.635	0.635	0.635	0.635
	level 2 (school)	0.0	0.0	0.0	0.0
Intracluster correlations (ICCs)	level 2 (school)	0.07	0.07	0.07	0.07
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15.7	3.5	21.3	6.5
Average year group size[17]		31.4	7	42.5	13
Number of year group	intervention	165	165	173	173
	control	165	165	173	173
	total	330	330	346	346
Number of schools	intervention	165	165	173	173
	control	165	165	173	173
	total	165	165	173	173
Number of pupils	intervention	2587	578	3676	1125
	control	2587	578	3676	1125
	total	5174	1156	7353	2249

[17] Average year group size refers to the number of pupils per school per year group. Given that 50% of pupils would be tested in maths, the average cluster size is half of the average year group size.

Table 9: Minimum detectable effect size for science

		Protocol		Randomisation	
		All pupils	FSM (primary analysis)	All pupils	FSM (primary analysis)
Minimum Detectable Effect Size (MDES)		0.14	0.18	0.14	0.16
Pre-test/ post-test correlations	level 1 (pupil)	0.645	0.645	0.645	0.645
Pre-test/ post-test correlations	level 2 (school)	0.0	0.0	0.0	0.0
Intracluster correlations (ICCs)	level 2 (school)	0.09	0.09	0.09	0.09
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15.7	3.5	21.3	6.5
Average year group size[18]		31.4	7	42.5	13
Number of year group	intervention	165	165	173	173
	control	165	165	173	173
	total	330	330	346	346
Number of schools	intervention	165	165	173	173
	control	165	165	173	173
	total	165	165	173	173
Number of pupils	intervention	2587	578	3676	1125
	control	2587	578	3676	1125
	total	5174	1156	7353	2249

The revised calculations (Randomisation columns in Table 8 and Table 9) are based on the number of schools retained at the randomisation stage of condition allocation (n= 173). The average cluster size 42.5 pupils per year group per school was calculated using the actual data at that time. The calculations use the same core assumptions as those conducted in the evaluation protocol. As before, we assume a pupil-level correlation between baseline and endline of 0.635[19] and the school-level intra-cluster correlation is assumed to be 0.07 for the

[18] Average year group size refers to the number of pupils per school per year group. Given that 50% of pupils would be tested in science, the average cluster size is half of the average year group size.

[19] The correlation between KS1 maths attainment and GL Progress Test in Maths result is estimated to be 0.76 based on FFT Education Datalab (2019) while the correlation between EYFSP overall point score and this Progress Test in Maths result is estimated to be 0.51 based on Roy et al. (2019). We use the average of these correlations for the pupil-level correlation between baseline and endline for the primary outcome measure.

primary outcome (maths).[20] We also make a conservative assumption that a school-level correlation between baseline and endline is 0.0. We use a Type I error rate of 0.05 and a Type II error rate of 0.20 (power of 0.80).

Using these revised sample sizes, we find that the randomisation sample sizes would yield an MDES[21] of 0.14 standard deviations for the primary analysis for maths attainment among KS2 pupils eligible for FSM. The MDES of 0.14 is lower than our initial expectation of 0.17 in the evaluation protocol. This indicates that the trial is so far on track, at least meeting the MDES assumptions outlined in the evaluation protocol, even if there might be further attrition by the time of the endline data collection.

Analysis

Primary outcome analysis

The evaluation of Stop and Think aims to estimate the impact of the programme on maths and science attainment, using an intention-to-treat (ITT) approach. The trial is designed as a two-armed cluster randomised control trial with pupils clustered within schools.

The primary outcome of interest is maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds (as defined as those who have been eligible for FSM at any point in the previous 6 years) (RQ1). As mentioned in the Primary outcome section, the primary outcome measure will be a standardised measure of pupils' mathematical skills and knowledge, the GL Progress Test in Maths (GL PTM) for pupils eligible for FSM. Based on this, the primary analysis will be a subgroup analysis including only Year 3 and Year 5 pupils eligible for FSM. As suggested by the EEF analysis guidance[22], the variable "EVERFSM_6_P_[term][yy]" from the National Pupil Database (NPD), which indicates if a pupil has been recorded as eligible for FSM at any time in the last 6 years, will be used to identify pupils coming from a disadvantaged background.

The primary outcome analysis will estimate the pooled effect of Stop and Think on maths attainment by including both Year 3 and Year 5 pupils. Note that the protocol proposed a IPD meta-analysis approach using a three-level model to estimate a single effect for the primary outcome. Yet we have modified the approach to use a simpler model that achieves same goal robustly. [23] The model will instead be a two-level linear regression, with pupils at level one and schools at level two. We are using a multilevel model to account for the fact that pupils are clustered within schools. The year group will be added as a covariate (i.e. a fixed effect) to account for a year group information.

Following Ashraf et al. (2021), we will scale the baseline measures to a unit variance of 1 per trial to eliminate heterogeneity between year groups. The scaling of raw measures in each

[20] As found in the efficacy trial. Roy, P. et al. (2019) *Stop and Think: Learning counterintuitive concepts. Evaluation report.* Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop_and_Think.pdf?v=1680163623

[21] MDES indicates the smallest effect size that an impact evaluation is able to detect for a given level statistical significance and power.

[22] EEF (2022) Statistical analysis guidance for EEF evaluations. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1679395501

[23] During the protocol writing stage, the IPD meta-analysis approach was considered for estimating a single overall effect for Year 3 and Year 5 pupils combined for the primary outcome. However, from further discussions and suggestions from one peer reviewer and the university of Durham, it was concluded that a simpler approach would be an appropriate model as the evaluation involves only two year groups and the endline outcome measures are similar. We follow a two-level model with year groups added as a fixed effect to estimate a pooled effect of the programme on the primary outcome.

year group will be completed separately by using mean and standard deviation of the scores within each year group.

The *age-standardised* PTM score from age-appropriate tests will be used as the dependent variable in this model, with a binary indicator of treatment allocation, *standardised* pre-trial test scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils), year group fixed effect and the randomisation strata as covariates.[24]

The basic form of the model for pupils eligible for FSM across both year groups is:

$$PTM_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 Stratification'_j + \beta_4 YearGroup_{ij} + u_j + e_{ij}$$

where pupils eligible for FSM (i) are clustered within schools (j). β_0 is an overall intercept, β_1 is a fixed gradient between the standardised post-test and pre-test scores and β_2 is the average effect of the intervention. 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 stratification variables used for randomisation and year group information will be included as fixed effects in this model.[25] School-level random effects will account for school-level variation in outcomes that is not explained by the fixed effects. In line with the EEF analysis guidance (EEF, 2022), other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 14 using the **mixed** command.[26]

The impact of the intervention will be expressed as a standardised effect size. 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 for FSM pupils, along with a summary of means and standard deviations of pre- and post-test scores.

Secondary outcome analyses

The secondary outcomes include maths attainment for all Year 3 and Year 5 pupils (RQ2), science attainment for all Year 3 and Year 5 pupils (RQ3), science attainment for Year 3 and Year 5 pupils from disadvantaged backgrounds as measured by FSM status (RQ4) and the raw scores of common misconceptions in maths (RQ5) and science (RQ6) among all Year 3 and Year 5 pupils. As mentioned in the Secondary outcome section, we will measure maths and science attainment following intervention delivery by administering age-specific GL Progress Tests and measure common misconceptions using age- and subject-specific tests being developed by NatCen and Oxford MeasurEd. Individual pupils will be randomly assigned to sit either maths or science tests, so that 50% of pupils in each year group will be tested in maths attainment and maths misconceptions and 50% in science attainment and science misconceptions. More details on the outcome measures are provided in the protocol (the Outcome measures section).

[24] We will use a variable named "KS1_MATH_OUTCOME" from the NPD as the baseline measure for Year 3 pupils. Following the efficacy trial, we will use average EYFSP point score, which will be formed by combining all 17 early learning goals, as the baseline measure for Year 5 pupils.

[25] Schools were stratified by class-form entry size and the school-level proportion of pupils eligible for FSM at any time during the past 6 academic years.

[26] As the baseline data will be supplied from NPD, the analysis will need to be conducted through the Office for National Statistics Secure Research Service (ONS SRS). Stata 14 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 https://www.ons.gov.uk/aboutus/whatwedo/statistics/requestingstatistics/approvedresearcherscheme [Assessed 22/06/2023]

Each model below for the secondary outcome analyses will include a binary indicator of treatment allocation, and the randomisation strata and year group variables as fixed effects. In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 14 using the **mixed** command.

Analysis for attainment in maths and science

For RQ2-RQ4, we will estimate the impact of the programme on attainment in science (amongst all pupils and amongst FSM pupils) and maths (amongst all pupils) using the same approach as outlined for the Primary outcome analysis above. For each secondary outcome, two-level model will be estimated for both year 3 and year 5 pupils combined, to reflect pupils (i.e. level 1) nested within schools (i.e. Level 2). The age-standardised PTM or PTS score will be used as the dependent variable in these models. These models will include standardised pre-test scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate[27].

The basic form of each model for pupils in both year groups is:

$$Outcome_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 Stratification'_j + \beta_4 YearGroup_{ij} + u_j + e_{ij}$$

where pupils eligible for FSM (i) are clustered within schools (j). β_0 is an overall intercept, β_1 is a fixed gradient between the standardised post-test and pre-test scores and β_2 is the average effect of the intervention. The term u_j is a school-level random effect and e_{ij} is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model.

The impact of the intervention will be expressed as a standardised effect size. See the Effect size calculation below for an explanation of how effect sizes will be calculated.

Analysis for misconception in maths and science

To measure the effects of Stop and Think on misconceptions in maths and science (RQ5-RQ6), we will estimate the effects in a similar way to how we estimate the impact of the programme on the primary outcome (RQ1). The model will follow an ITT approach and will include baseline measures of attainment in maths. As mentioned, the misconception outcome measures will be raw scores representing the number of times the learner fell into a common misconception. The total number of misconceptions will be analysed using a multilevel Poisson regression. The basic form of the model for both year groups combined is as follows:

$$Outcomes\_Misconception_{ij} = exp(\beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 Stratification'_{ij} + \beta_4 YearGroup_{ij} + u_j + e_{ij})$$

where pupils (i) are clustered within schools (j). $\exp(\beta_0)$ is an overall intercept and $\exp(\beta_2)$ is the average effect of the intervention. The term $\exp(u_j)$ is a school-level random effect and

[27] We will use a variable named "KS1_MATH_OUTCOME" from the NPD as the baseline measure for Year 3 pupils. Following the efficacy trial, we will use average EYFSP point score, which will be formed by combining all 17 early learning goals, as the baseline measure for Year 5 pupils.

$\exp(e_{ij})$ is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model.

Additional analyses

Sensitivity analysis for primary outcome

As for the primary outcome, the approach used in the efficacy trial will also be used as a sensitivity analysis. For maths attainment outcome, a two-level fixed effects model will be built for each year group with the raw PTM score as the dependent variable. The model will reflect the structure of the data with pupils nested within schools. Each model will include standardised pre-trial test score (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate, as outlined above

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

$$PTM_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + u_j + e_{ij}$$

where pupils (i) are clustered within schools (j). The intervention effect is estimated by β_2. The term u_j is a school-level random effect, and e_{ij} the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. For these measures we will report confidence intervals at 95% level and the effect size using Hedges' g formula as described in the later section.

To estimate a single effect size for both Year 3 and Year 5 pupils for the maths outcome from separate models for each year group, the mean of the two resulting effect sizes will be taken to calculate a single effect size that is comparable with findings from other studies, including the Stop and Think efficacy trial. The variance of the combined effect size will be estimated using the formula in Borenstein, Hedges, Higgins, & Rothstein, (2009, p. 218), following the EEF statistical guidance.[28]

To calculate a precise estimate of the overall effect size for both Year 3 and Year 5, we will assign the weight to each effect size Y_i using the formula as follows:

$$W_i = \frac{1}{V_{Y_i}}$$

where V_{Y_i} represents the within-model variance for model (i). Given that we will have only two results (Year 3 and Year 5) for the primary outcome, the weighted mean (M) can be computed as

$$M = \frac{W_1 Y_1 + W_2 Y_2}{W_1 + W_2}$$

The variance of the summary effect is then obtained as

[28] Borenstein, M., Hedges, L. V., Higgins, J. P T., & Rothstein, H. R. (2009). Introduction to meta-analysis. London: Wiley, pp. 63-67; pp.217-223.

$$V_M = \frac{1}{W_1 + W_2}$$

Further sensitivity analysis

As further sensitivity analysis, an alternative model will be estimated to assess whether the findings for the primary analysis are robust to different model specifications. We will estimate the impact of the programme on the sample as a whole with a two-level model reflecting pupils nested within schools, which includes an interaction term between the treatment status and a dummy variable indicating FSM status. Furthermore, if differential loss to follow-up creates an imbalance between trial groups or if attrition is high, the sensitivity of the estimated effect will be assessed by approximating missing outcomes using multiple imputation. The model will include standardised pre-trial test score (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate. The basic form of the model for pupils i is:

$$PTM_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 FSM_{ij} + \beta_4 Intervention_j FSM_{ij} + \beta_5 Stratification'_j + \beta_6 YearGroup_{ij} + u_j + e_{ij}$$

where pupils eligible for FSM (i) are clustered within schools (j). β_4 is the attainment gap (i.e. difference in average effect of the intervention between FSM pupils and their peers). β_2 is the impact of the intervention on non-FSM pupils and the impact of the intervention on FSM pupils is $\beta_2 + \beta_4$.

The term u_j is a school-level random effect and e_{ij} is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. The stratification and year group variables will be included as fixed effects in this model while the school-level random effects will control for other observed and unobserved school-level characteristics.[29] In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 14 using the **mixed** command.

The effect of the intervention on attainment gap will be estimated. See the Effect size calculation section below for an explanation of how effect sizes will be calculated.
In case of high attrition and missing data imputation for the primary outcome, we will conduct further sensitivity analysis to assess whether results from the imputed data differ from the complete data, following EEF analysis guidance (2022). Details on missing data imputation are covered in the Missing data section.

Mediation analysis

Mediation analysis is used to explore mechanisms by which an intervention affects the outcomes of interest. For the Stop and Think evaluation, one of the proposed mechanisms by which the intervention affects maths attainment among Year 3 and Year 5 pupils eligible for FSM is by improving curriculum-appropriate maths misconception. This is reflected in the logic model presented in the study protocol. [30]

[29] Schools were stratified by class-form entry size and the school-level proportion of pupils eligible for FSM at any time during the past 6 academic years.
[30] The logic model is available as Figure 2 in the Stop and Think evaluation report which is available athttps://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1648826229

The maths misconception. test will be administered at endline and the score from this test will be used for the mediation analysis.

An exploratory analysis will be conducted to decompose the intention-to-treat estimate into an indirect effect (i.e. effect of the intervention that can be attributed to changes in the common misconceptions in maths) and a direct effect (i.e. effect of the intervention that cannot be attributed to changes in the common misconceptions in maths). The assumed causal model for this analysis is shown in Figure 2 below.

Mediation analysis will be conducted to understand whether the effect of the Stop and Think programme on pupils' attainment is partially or totally mediated by changes in the common misconceptions in maths. Our hypothesis is that the effect will be at least partially mediated, but we do not have an expectation of the magnitude of this effect. The mediation analysis will follow the steps below:

- Step one (Path a): regress pupils' common misconceptions in maths on the Stop and Think programme.

$$Maths\_Misconceptions_{ij} = \beta_0 + \beta_1 Intervention_j + \beta_2 YearGroup_{ij} + \beta_3 Stratification'_j + u_j + e_{ij}$$

The slope β_1 tells us how much pupils' common misconceptions change between those who were part of Stop and Think programme and those who were not.

- Step two (Path c): regress pupils' attainment in maths on Stop and Think programme and pupils' common misconceptions in maths.

$$PTM_{ij} = \alpha_0 + \alpha_1 Intervention_j + \alpha_2 YearGroup_{ij} + \alpha_3 Stratification'_j + \alpha_4 Maths\_Misconceptions_{ij} + u'_j + e'_{ij}$$

The slope α_1 provides average direct effect, slope β_1 shows how much Stop and Think programme shifts pupils' common misconceptions in maths, α_4 tells us how much pupils' maths attainment changes for a unit increase in pupils common misconceptions. $\beta_1\alpha_4$ gives us Average Causal Mediation Effect (ACME).

- Step three: we will estimate the Average Casual Mediation Effect (ACME) using the *mediation* package in R (Imai et. al., 2010). We will also report the *proportion mediated* estimate, that is, the magnitude of the mediated effect relative to the total effect. We will also report 95% confidence intervals using bootstrapping with 1000 simulations.

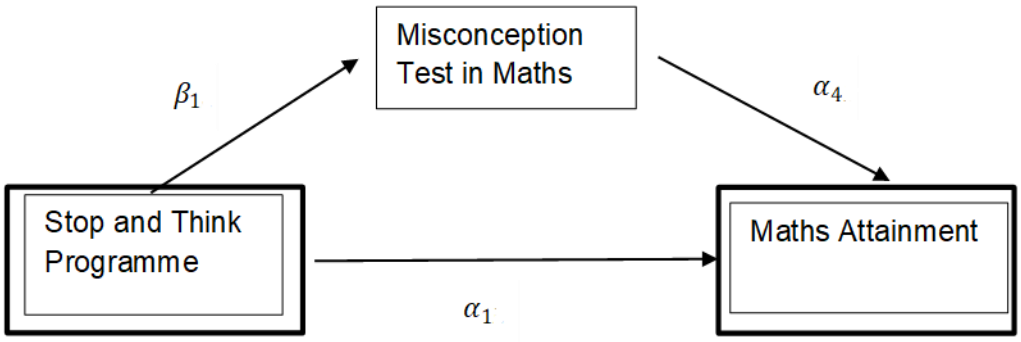
$$\textbf{Proportion mediated} = \frac{\textbf{Mediated effect}}{\textbf{Total effect}} = \frac{\boldsymbol{\beta_1\alpha_4}}{\boldsymbol{\beta_1\alpha_4 + \alpha_1}}$$

For all steps, we will present the unstandardised model coefficients, p-values, and 95% confidence intervals. The primary effect size that we will interpret is the proportion mediated and its confidence interval, since – given that it is a proportion – this is the most straightforward to understand.

Figure 2 below shows the casual mediation model. α_1 shows the average direct effect of Stop and Think programme on maths attainment**.** β_1 tells us how much pupils' common misconceptions change between those who were part of Stop and Think programme and those who were not. α_4 tells us how much pupils' maths attainment changes for a unit

increase in pupils common misconceptions. $\beta_1\alpha_4$ is the indirect effect of Stop and Think programme on maths attainment.

Figure 2 Casual Mediation Model



Imbalance at baseline

To check for, and monitor, imbalance at baseline following baseline assessment, we will undertake descriptive analysis at school and pupil level.

We will firstly assess imbalance at baseline by condition allocation[31], covering:

- Class-form entry size
- School-level proportion of pupils eligible for FSM at any time during the past 6 academic years

By year group, the baseline comparisons between treatment and control groups at pupil level will cover:

- Ever received FSM
- KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils

Categorical variables will be explored by conducting cross-tabulations, including counts and percentages in each category. Continuous variables will be summarised with descriptive statistics (n, mean, standard deviation, range, median and effect sizes) by condition/group allocation. We will report standardised mean differences in baseline characteristics as Hedges' *g* effect sizes. An effect size of greater than 0.05 will be considered as an indication of possible imbalance. Note that the analyses will be performed through ONS SRS workspace, the outputs will thus have to follow SRS rules on statistical disclosure control.

If imbalances are indicated, a sensitivity analysis will be estimated. This model will include the unbalanced variables (i.e., where Hedges' *g* is greater than 0.05) in addition to those in the main model will be estimated as a sensitivity analysis.

Missing data

As a first step, we will explore the extent of missing data on the outcome and pre-treatment covariates descriptively, with cross-tabulations, including counts and percentages in each category.

[31] As mentioned, randomisation was conducted by condition 1 (i.e., Year 3 allocated to treatment and Year 5 allocated to control) and condition 2 (i.e. Year 5 allocated to treatment and Year 3 allocated to control)

To better understand the pattern of missing data, we will explore the extent of missingness, and whether there is a pattern in missingness. A 'drop-out' model will be estimated using a logistic regression to assess if there are patterns to missing data. The outcome will be binary, reflecting whether the primary outcome data, and any covariates from the primary analysis are missing for each individual at follow-up. This model will include all covariates outlined in the 'Imbalance at baseline ' section, in addition to a random effect for schools. Missing data for these covariates will be coded up as separate binary variables in the model. The 'drop-out' model will be estimated using the *melogit* command in Stata.

We will follow the protocol for missing data suggested by the EEF (see EEF, 2022). For less than 5% missingness overall from randomisation to final analysis, a complete-case analysis will be employed. For more than 5% missing data overall from baseline assessment to final analysis, our approach will depend on pattern of missingness. If the pattern of missingness may be unrelated to the treatment effect (e.g., absence due to pupil illness, staff changes, or other factors that affected testing but are not related to Stop and Think), then missing data will be assumed MCAR and we will continue with a complete case analysis. If data is missing in a way that is correlated with observable variables, the primary analysis will be re-estimated through Multiple Imputation (MI) using Chained Equations (MICE). To do so, the probability that the outcome measure was missing will be modelled using a multilevel logistic model that includes the covariates involved in the primary analysis. The covariates would thus include baseline attainments, intervention allocation, randomisation stratification, FSM status, and so on. The significant variables would then be used for a MI process. The minimum number of imputed datasets will depend on the fraction of missing information, as suggested by Graham et. al. (2007). The imputed datasets will be used to replicate the main analyses and we will compare the results with the complete data analysis as part of sensitivity analyses.

If the pattern of missingness depends on an unobserved variable, even after considering all the information in the observed variables, we will consider the missing observations are missing not at random (MNAR) and follow EEF guidance (2022) to carry out a weighting approach after MI, as suggested by Carpenter et al. (2007). Note that missing data analysis will only be possible in cases where we have data from the NPD (i.e. FSM and baseline attainments).

Multiple imputation will be conducted using the mi suite of commands in Stata 14. The imputation method will depend on the number and types of significant variables that will be used for a MI process (StataCorp, 2013, pp.114).

Compliance

The Complier Average Causal Effect (CACE)[32] will be estimated to show the impact of Stop and Think on the primary outcome (maths attainment among KS2 pupils eligible for FSM) compared to individuals in the control group, taking into account level of compliance with Stop and Think.

Data for our compliance analyses will be collected during the implementation through the computer-assisted programme that delivers Stop and Think. This will be used to measure if and how fully the intervention has been delivered to classes.[33] Specifically, the computer-assisted programme will count the number of completed sessions delivered to each class

[32] Corresponding to the average effect of the intervention for those pupils who have complied with the programme.
[33] Additional data will be available from the developer on the average amount of time per session, the spacing of sessions and the number of structured practice activities completed

(from 0 to 30). For completed sessions, we mean sessions started in which at least one question was answered. Following the approach used in the efficacy trial, we will use the number of completed sessions delivered to each class as a continuous measure of compliance. We will assume one-sided non-compliance in our analysis as we assume that none of the pupils in the control group can be exposed to these sessions since the control group receive a business-as-usual teaching approach and do not have access to Stop and Think.

Although the measures of compliance are at class level, the unit of analysis will be pupils. Further, considering that schools may have unobserved characteristics influencing both the compliance with the intervention and the primary outcome, we will estimate the CACE using a two-stage least square (2SLS) model (Angrist and Imbens, 1995) with the treatment allocation as the instrumental variable (IV) for the compliance measure.

The first stage of the model will be compliance regressed on all covariates that are used in the main primary outcome model and in addition, will include, as an IV, a binary variable that indicates a pupil's pre-intervention treatment allocation. The first stage equation estimate is as follows:

$$Comply_j = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 YearGroup_{ij} + \beta_4 Stratification'_j + e_{ij}$$

The second stage of the model will regress the primary outcomes on the covariates used in the main models and will also include a covariate representing the pupil's estimated level of compliance from the first stage of model and an interaction term between the estimated compliance and the pupil's pre-intervention treatment allocation. The estimation of the second stage equation is as follows:

$$PTM_{ij} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 YearGroup_{ij} + \beta_3 Stratification'_j + \beta_4 \widehat{Comply}_j + e_{ij}$$

The coefficient (β_4) 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.

Note that we will not block use of software. Instead, we conduct two sets of compliance analysis. We will first look at the number of completed session delivered until the intervention end date. We will then provide an additional robustness check where we redo the compliance analysis using the total number of completed sessions delivered up until the endline testing rather than intervention end date. This means we will have two compliance measures: a) compliance truncated to the intervention end date and b) compliance untruncated until the endline testing.

IV regression will be conducted in Stata 14, using the ***ivregress*** command and the ***cluster*** option to control for clustering on schools.

Intra-cluster correlations (ICCs)

The intra-cluster correlations (ICCs) will be estimated directly from the primary analysis model, using the variance estimates for each level of clustering. The ICC for schools $\boldsymbol{\rho_S}$ will be estimated with the post-estimation command **estat icc** in Stata 14, using the following formula based on Hedges (2011):

$$\boldsymbol{\rho_S = \frac{\sigma^2_{BS}}{\sigma^2_{BS} + \sigma^2_{WS}} = \frac{\sigma^2_{BS}}{\sigma^2_{WT}}}$$

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

Effect size calculation

Effects size calculation for primary and secondary outcome analyses

We will use the effect sizes (ES) for cluster-randomised trials, as adapted from Hedges (2007):

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

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

From the primary outcome model, we will take each group's adjusted mean and variance to calculate the effect size. The variance will be the total variance (across both pupil and schools, 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, will also be reported. The ES will be estimated using the eefanalytics Stata package.[34]

Effects size calculation for the sensitivity analysis for primary outcome

As mentioned, we will report the mean effect size of the two-level model for Year 3 and Year 5 pupils, using the approach followed in the Stop and Think efficacy trial.[35] We will use the effect sizes (ES) for cluster-randomised trials, as adapted from Hedges (2007):

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

Where $(\overline{Y_T} - \overline{Y_C})_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics, while $\sqrt{\sigma_u^2 + \sigma_e^2}$ is an estimate of the population standard deviation. σ_u^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 will be the total variance of both groups (across both pupil and schools, 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, will also be reported. Y_3 and Y_5 are the effects sizes and V_3 and V_5 are the variances for the Year 3 and Year 5 models, respectively.

Roy et al. (2019) followed the method described by Borenstein et al. (2009, p.218) to obtain the combined effect size.

---

[34] For more information see EEFANALYTICS: Stata module for Evaluating Educational Interventions using Randomised Controlled Trial Designs (repec.org)

[35] Roy, P. et al. (2019) *Stop and Think: Learning counterintuitive concepts. Evaluation report.* Available at: https://www.nfer.ac.uk/media/3703/learning_counterintuitive_concepts_evaluation_report_-final.pdf.

To obtain the combined effect size, we will first calculate the weights assigned in each model:

$$W_3 = \left(\frac{1}{V_3}\right) and\ W_5 = \left(\frac{1}{V_5}\right)$$

Where, $V_3\ and\ V_5$ are variances for the Year 3 and Year 5 models, respectively.

The combined effect size will then be calculated as:

$$Y_c = \frac{(Y_3 * W_3) + (Y_5 * W_5)}{(W_3 + W_5)}$$

Lastly, the combined variance will be calculated as:

$$V_c = \frac{1}{(W_3 + W_5)}$$

Effects size calculation for further sensitivity analysis

We will use the effect sizes (ES) for cluster-randomised trials, as adapted from Hedges (2007):

$$ES = \frac{(AttainmentGap)_{adjusted}}{\sqrt{\sigma_u^2 + \sigma_e^2}}$$

Where $(AttainmentGap)_{adjusted}$ (i.e. β_4 as per in the model above) the difference in average effect of the intervention between FSM pupils and their peers adjusted for baseline characteristics, while $\sqrt{\sigma_u^2 + \sigma_e^2}$ is an estimate of the population standard deviation. σ_u^2 is the variance of school level intercept and σ_e^2 is variance of residuals. A 95% CI for the ES, that takes into account the clustering, will also be reported.

References

Angrist, J. D. and Imbens, G. W. (1995) 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association*. 90 (430), pp. 431–442. DOI 10.1080/01621459.1995.10476535

Ashraf et al. (2021). *Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 – 2019.* EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-syntheses/Individual-participant-data-meta-analysis-of-the-impact-of-EEF-trials-on-the-educational-attainment-of-pupils-on-Free-School-Meals.pdf

Azubuike, O. B., Moore, R. and Iyer, P. (2017) *The design and development of cross-county Maths and English tests in Ethiopia, India and Vietnam.* Technical Note 39. Oxford: Young Lives.

Borenstein, M., Hedges, L. V., Higgins, J. P T., & Rothstein, H. R. (2009). Introduction to meta-analysis. London: Wiley, pp. 21-32.

Carpenter, J.R., Kenward, M.G. and 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.

Carril, A. (2017). Dealing with misfits in random treatment assignment. *The Stata Journal*, *17*(3), 652-667.

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 (2022) 'Statistical Analysis Guidance for EEF Evaluations'. Available at https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1679395501

From FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One. London: FFT E.D. Available at https://ffteducationdatalab.org.uk/2019/10/how-attainment-gaps-emerge-from-foundation-stage-to-key-stage-4-part-one/

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention science*, *8*, 206-213.

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

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25, 51–71.

Larry V. Hedges, 'Effect Sizes in Three-Level Cluster-Randomized Experiments', *Journal of Educational and Behavioral Statistics* 36, no. 3 (1 June 2011): 360, doi:10.3102/1076998610376617

Roy, P. et al. (2019) Stop and Think: Learning counterintuitive concepts. Evaluation report. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/projects/Stop_and_Think.pdf?v=1680093567.

StataCorp, L. P. (2013). Stata multiple-imputation reference manual. Available at https://www.stata.com/manuals/mi.pdf.