Education
Endowment
Foundation

STOP AND THINK
LEARNING COUNTERINTUITIVE CONCEPTS

Evaluation Report

March 2025

Helena Takala, Tien-Li Kuo, Enes Duysak, Sehaj Bhatti, Ekaterina Stoilova, Alina Fletcher, Nicky McGuinness, Mary McKaskill, Andi Fugard

National Centre
for Social Research

The Education Endowment Foundation (EEF) is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries, and colleges to improve teaching and learning for 2 to 19-year-olds through better use of evidence.

We do this by:

- **Summarising evidence.** Reviewing the best available evidence on teaching and learning and presenting in an accessible way.
- **Finding new evidence.** Funding independent evaluations of programmes and approaches that aim to raise the attainment of children and young people from socioeconomically disadvantaged backgrounds.
- **Putting evidence to use.** Supporting education practitioners, as well as policymakers and other organisations, to use evidence in ways that improve teaching and learning.

We were set-up in 2011 by the Sutton Trust partnership with Impetus with a founding £125m grant from the Department for Education. In 2022, we were re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

For more information about the EEF or this report please contact:

Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator

The project was independently evaluated by a team from NatCen Social Research: Helena Takala, Mary McKaskill, Enes Duysak, Sehaj Bhatti, Tien-Li Kuo, Ekaterina Stoilova, Alina Fletcher, Nicky McGuinness, Andi Fugard, Maha Basharat, Isabel Taylor, Padmini Iyer, and Lydia Marshall.

Mary McKaskill was the principal investigator until the analysis and the reporting stage, including for endline testing. Helena Takala was the principal investigator during analysis and reporting.

Contact details

Helena Takala
NatCen Social Research
35 Northampton Square, London EC1V 0AX

Email: **Helena.Takala@natcen.ac.uk**

Acknowledgements

We would like to thank Denis Mareschal (Birkbeck, University of London), Kathryn Atherton, and Callum O'Mahony (Behavioural Insights Team) for their collaboration on the project. They have been a pleasure to work with throughout and always generous with their time and attention. We are grateful to the teachers, senior leaders, and primary school pupils who contributed their time and views for the study and enabled endline data collection during a busy time for schools. From NatCen, we are indebted to Emma Fenn, Sandie Napier, Peter Baker, and Bryan Mason for planning and managing large-scale data collection and transcription and for the field interviewers who invigilated testing in schools across England. Finally, we want to thank colleagues at the EEF, particularly Faizaan Sami and Jamila Boughelaf, for their support throughout.

This work was undertaken in the Office for National Statistics' (ONS) Secure Research Service. Using data from the ONS and other owners and does not imply the endorsement of the ONS or other data owners.

Executive summary

The project

Stop and Think is a computer programme, developed by academics at Birkbeck, University of London and IOE (UCL's Faculty of Education and Society), which aims to improve pupils' ability to adapt to counterintuitive concepts in maths and science. It trains pupils to inhibit their initial, intuitive responses to questions surrounding such concepts and give slower, more reflective answers instead. The programme targets Year 3 (ages seven to eight) and Year 5 (ages nine to ten) pupils, with content aligned to the national curriculum.

The intervention involved three 12-minute sessions being delivered each week over a ten-week period – a total of 30 sessions. Sessions include games with multiple-choice answers that address common misconceptions. Teachers delivered the intervention at the start of maths and/or science lessons using a computer and a projector or an interactive whiteboard. The programme is a whole-class activity with pupils working through problems together.

This effectiveness trial involved 173 primary schools across England – a total of 14,718 pupils. Schools were randomly assigned to have either Year 3 or Year 5 pupils receive Stop and Think while the other year group continued with 'business as usual'. The evaluation used a two-arm cluster randomised controlled trial design to assess the impact on pupil maths and science attainment. The primary outcome was maths attainment among Year 3 and Year 5 pupils eligible for free school meals (FSM), measured using progress tests by GL Assessment. Secondary outcomes included maths and science attainment for all pupils, and the prevalence of common misconceptions. The impact evaluation was accompanied by an integrated implementation and process evaluation (IPE) design. The trial ran from October 2022 to September 2023, having faced delays due to the COVID-19 pandemic after its initial start in July 2020.

Key conclusions
1. Pupils eligible for FSM receiving Stop and Think made no additional months' progress in maths attainment compared to FSM pupils receiving teaching as usual. This result has a moderate to high security rating. All pupils receiving Stop and Think made no additional months' progress in maths attainment compared to pupils receiving teaching as usual.
2. All pupils receiving Stop and Think made two additional months' progress in science attainment compared to pupils receiving teaching as usual. FSM pupils receiving Stop and Think made one additional month's progress in science attainment, compared to FSM pupils receiving teaching as usual. Further analysis suggests that the impact on FSM-eligible pupils was similar to the impact for all pupils in science.
3. The Stop and Think programme largely took place as intended with most participating teachers delivering the 30 intervention sessions. However, due to scheduling and staffing issues, some teachers could not always follow the model of three sessions per week at the start of maths and/or science lessons within the ten-week period.
4. Evidence to support the short-term outcomes of reduced impulsive responding and improved attainment on science and maths misconceptions tests is inconclusive due to the low reliability of the assessment methods used. No link was found between socio-economic status (measured by FSM), increased participation, or compliance with Stop and Think and improved maths attainment.
5. Some teachers and pupils perceived the maths content to be easier in comparison to the science content. There is evidence to suggest the programme also showed only minor differences from usual maths and science teaching practice. This perception may have been because they saw the programme as an extension of their usual science and maths teaching rather than a programme for enhancing pupils' inhibitory control.

EEF security rating

These findings have a moderate to high security rating. This was an effectiveness trial, which tested whether the intervention worked under everyday conditions in a large number of schools. The trial was a well-designed, well-powered two-arm cluster randomised controlled trial; 22.48% of the pupils who started the trial were not included in the final analysis because of school dropout, pupil absence on testing days, and non-consent to process pupil data.

Additional findings

Pupils eligible for FSM in Stop and Think schools made, on average, no additional months' progress in maths compared to those in the control group. This is our best estimate of impact, which has a moderate to high security rating. As with any study, there is always some uncertainty around the result: the possible impact of this programme also include small negative effects of one month of less progress to positive effects of up to two months of additional progress.

Pupils in Stop and Think schools made, on average, two additional months' progress in science compared to those in the control group, however, the possible impact of this programme also includes positive effects of only one month of additional progress. Participating FSM-eligible pupils achieved, on average, one additional month of progress in science compared to their counterparts receiving teaching as usual, however, the possible impact range includes no additional progress to positive effects of up to two months.

The IPE findings provide context for the differing impacts in maths and science. Some teachers and pupils found the maths content 'too easy', possibly due to its later delivery in the school year and the comparatively greater emphasis on maths in the Key Stage 2 curriculum. In contrast, the science content was seen as higher quality and more 'surprising' to pupils.

The programme aims to improve inhibitory control, not subject knowledge. Pupils' familiarity with the maths content may have limited their opportunities to effectively practice inhibitory control. While most schools completed the required number of sessions, the varying effectiveness of sessions may have resulted in insufficient practice of 'stop and think' skills.

The differing impacts on science and maths in both EEF trials suggest stronger subject-specific links than the logic model anticipated. This may be due to the greater emphasis on maths in the Key Stage 2 curriculum or fundamental differences in misconceptions between the subjects.

Teachers reported that the programme led to increased reflection on their practice, particularly in science, and improved classroom culture by reducing stigma around different processing times for pupils, including those with SEND and EAL.

Cost

The average cost of Stop and Think for one class was around £356 or £14.24 per pupil per year when averaged over three years.

Impact

Figure 2: Summary of impact on primary and secondary outcomes

Outcome/ Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	P Value	EEF cost rating
Maths, FSM pupils	0.04 (-0.06, 0.13)	0	🔒🔒🔒🔓🔓	1841	0.423	**£** £ £ £ £
Maths, all pupils	0.04 (-0.01, 0.09)	0		6044	0.174	**£** £ £ £ £
Science, FSM pupils	0.08 (-0.01, 0.17)	1		1819	0.112	**£** £ £ £ £
Science, all pupils	0.11 (0.05, 0.16)	2		6033	0.001	**£** £ £ £ £

Introduction

Background

Policy context

Existing evidence indicates that there is room for improvement in the maths and science skills of young people in England compared to international standards. The Programme for International Student Assessment (PISA) results in 2022 showed that 15-year-olds in England ranked 13th in science and 11th in maths, out of 81 participating countries. England's maths results reduced by 12 points compared with the last PISA round in 2018, while attainment in science, which has been in long-term decline and decreasing by five points in each iteration, went down a further four points (OECD, 2022). A 2016 study found that over a quarter of young people aged 16 to 19 in England had low numeracy skills, with England 22nd out of 23 countries (Kuczera et al., 2016). In a 2022 study, employers felt labour market entrants are not properly prepared for the workforce, with the U.K. again comparing poorly against other countries in this respect (UK Commission for Employment and Skills, 2022). There is also evidence of intergenerational effects, with poor parental attainment in maths and science reflected in young people's educational outcomes (OECD, 2017; Kuczera et al., 2016).

The U.K. government has identified improving science, technology, engineering and maths (STEM) skills as key to improving the international competitiveness of the U.K. economy (Department for Business, Energy and Industrial Strategy, 2024). This is particularly important in light of global technological advances; the automation of increasingly sophisticated tasks through artificial intelligence also means that required skills in the labour market are rapidly changing (Department for Science, Innovation and Technology, 2024). To meet these needs, successive U.K. governments have introduced measures to improve attainment in STEM subjects and to increase their take-up at Key Stages (KS) 4 and 5 (DfE, 2020). Research has shown that young people's choices and decisions about whether to take up STEM subjects at KS4 and KS5 are formed at primary school (Archer and Tomei, 2013). 'Appropriate, accurate and inspiring' STEM education in primary schools therefore plays a key role in later interest and engagement in STEM subjects and careers (Morgan et al., 2016).

Children's ability to learn maths and science concepts may be limited by their inability to inhibit perceptual evidence (what they see, feel, or hear) or their pre-existing beliefs (Vosniadou et al., 2018; Wilkinson et al., 2019). Compared to other subjects, there are many counterintuitive concepts in maths and science resulting in common mistakes because children tend to answer with an intuitive response (Babai et al., 2015). For example, in science, when children are taught that the world is round, there is no direct visual evidence to support this idea, as the horizon looks flat. Even after learning that the world is round, children may respond incorrectly when asked about the shape of the world because of their limited ability to inhibit their initial response. In maths, when learning about fractions, children may think that one quarter (¼) is larger than one half (½) because their initial response is to assess the denominators based on their knowledge of whole numbers.

Counterintuitive concepts are often the basis for common misconceptions in maths and science (Allen, 2014). To learn new concepts in these subjects, pupils must be able to inhibit prior contradictory knowledge and misconceptions (NFER, 2016). Existing evidence suggests this skill varies between pupils, with variation evident from an earlier age and weaker control skills among pupils from socio-economically disadvantaged backgrounds compared to their more advantaged peers (NFER, 2016).

Existing evidence for Stop and Think

Stop and Think is a computer programme that aims to raise KS2 maths and science attainment by improving pupils' ability to respond to and learn counterintuitive concepts. It was trialled for efficacy by the National Foundation for Educational Research (NFER) between 2015 and 2018. The efficacy trial, which was co-funded by the Wellcome Trust, was awarded high security (four 'padlocks') under EEF's Classification of the Security of Findings (EEF, 2019).

The efficacy trial had a within-school design with randomisation at the year-group level and was conducted in 89 schools in England. Year groups (Year 3 and Year 5) in each school were randomised to either take part in the intervention or to one of two control groups. The first control group received 'teaching as usual' (continuing with normal classroom practice), while the second group was an active control (receiving a computer programme to support social/emotional skills). By including an active control, the efficacy trial was able to measure the specific effects of Stop and Think beyond engagement and motivation caused solely by the novelty of playing a computer game.

The joint primary outcomes for the efficacy trial were the combined effect size (across Year 3 and Year 5) in maths (GL Progress Test in Maths) and combined effect size in science (GL Progress Test in Science). The trial also looked at a general measure of inhibitory control as a secondary outcome (Carlson et al., 2002).

The trial found that, on average, the intervention group made the equivalent of one additional month of progress in maths and two additional months' progress in science compared to both control groups. However, the evaluators reported that the effect for maths was not statistically significant (Roy et al., 2019). Pupils who received Stop and Think also made more progress than pupils in the active control group. Similar to pupils in the treatment group, pupils in the active control group received a computer-based learning programme but which had an unrelated focus on supporting social and emotional skills. The evaluators reported that these results were statistically significant for both maths and science. These results demonstrate that Stop and Think had an impact on pupils' maths and science attainment over and above a similar computer programme. The NFER, therefore, recommended that a subsequent effectiveness trial did not need to include an active control group.

The efficacy trial did not find an impact on the secondary intermediary outcome of pupil inhibitory control, which is considered important in helping pupils develop subject-specific reasoning skills (Roy et al., 2019). The post-intervention test used for this intermediate outcome in the efficacy trial required pupils to respond to as many questions as possible within a set time-limit. As such, its instructions were not in line with the main aim of the Stop and Think programme and its set-up may have actively discouraged pupils from pausing to consider the question before they answer. This focus on speed of response may have impacted these findings.

The efficacy trial was not powered to measure an effect for free school meal (FSM) pupils and the evaluators reported that the effects were not statistically significant. However, post-hoc sensitivity analysis carried out by academics at Durham University showed some initially promising results for FSM pupils: on average, these pupils made additional progress compared to the control group in (a) Year 3 and Year 5 maths and (b) Year 5 science (Durham University, 2020). Its analysis suggested a statistically significant effect of the programme on maths attainment among FSM pupils that was larger than the estimated effect size on science attainment or maths attainment among all pupils (Roy et al., 2019). NFER, therefore, recommended that a subsequent effectiveness trial should further explore the impact on pupils receiving FSM.

Intervention

Stop and Think is a computer 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 aim of the programme is to teach reflective skills to the pupils: it is not focused on teaching them maths and science content. The programme content includes a series of sessions. Each session includes games made up of science and maths questions and multiple-choice answers that include distractors demonstrating common misconceptions. The session topics are aligned to the maths and science curriculum in Years 3 and 5.

Intervention delivery

The Stop and Think programme was developed by academics at Birkbeck, University of London and the IOE, UCL's Faculty of Education and Society. In this evaluation, the intervention was coordinated and delivered by the Behavioural Insights Team (BIT). This was different from the efficacy trial where Birkbeck had developed as well as delivered the intervention. Within participating primary schools in England, the intervention was delivered by Year 3 and Year 5 teachers. Schools were asked to identify a school lead who was responsible for coordinating Stop and Think delivery and evaluation activities.

Intervention delivery started in February 2023 and lasted until May 2023. This was a change from the efficacy trial when classroom delivery took place slightly earlier in the academic year (November to March). The delivery team made this change in order to have sufficient time to recruit and train the larger numbers of schools ahead of the intervention period. From a curriculum perspective, this meant that pupils would have encountered more of the content covered in the programme by the time delivery started for this trial.

During the delivery period, schools were expected to deliver a total of 30 Stop and Think sessions, three times per week over a ten-week period. The first was an introductory session that did not involve any science or maths topic. This was followed by 29 sessions of the game before a science or maths lesson. Each session lasted around 12 minutes (including six minutes of maths content and six minutes of science content in each session). The dosage was unchanged from the efficacy trial. The intervention was delivered in classrooms at participating schools at the start of maths and/or science lessons.

Before the intervention started, a research colleague from BIT visited the schools in October 2022 to January 2023 to train treatment group teachers in how to deliver the intervention. They registered classes on the Stop and Think software, took teachers through one of the 30 sessions to show them how to navigate the software, and delivered a short training session on how to deliver sessions. Teachers were informed that they would need a computer and a projector or an interactive whiteboard to deliver the sessions: this was a prerequisite to taking part in the intervention.

Teachers were provided with an initial information session, login details, a handbook (including frequently asked questions), and a briefing video. Ongoing support was available from BIT by email and phone if teachers wished to access this. However, unlike the efficacy trial, support was not proactively offered by the delivery team based on recommendations from NatCen to replicate more 'real world' conditions during the effectiveness trial.

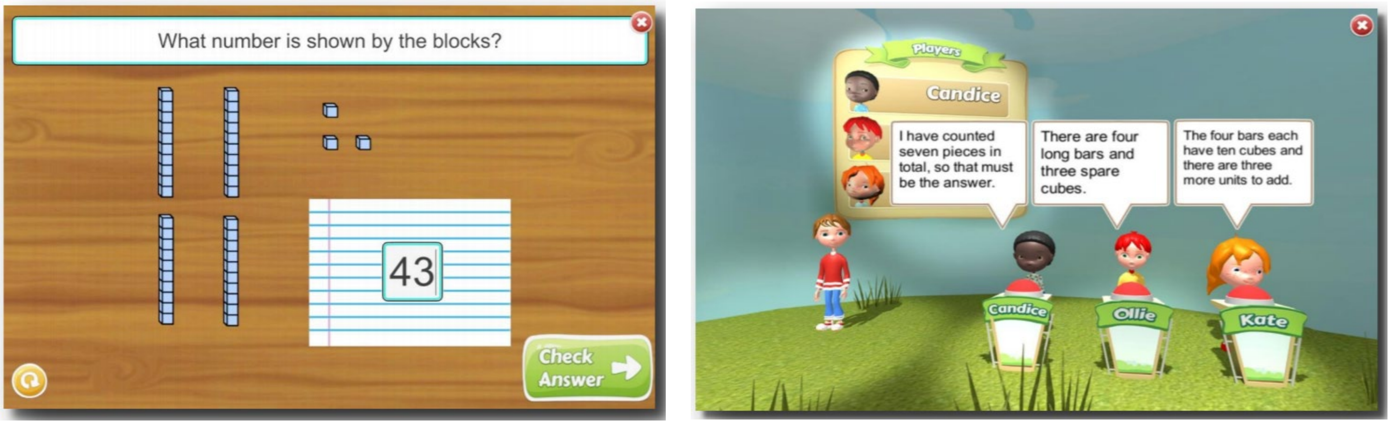
Intervention content

Stop and Think underwent some modifications and development which was co-funded by the Wellcome Trust after the efficacy trial (between September 2020 and March 2022). This included minor changes to the content and software, which were based on feedback from the efficacy trial, a design workshop with game designers, teachers and the unLocke team,[1] and a focus group with pupils to appraise the design of the software. Following modifications based on these consultations, Birkbeck ran a small-scale validation study in three schools to identify any remaining software issues.

Stop and Think uses a question-and-answer format. The session topics are aligned to the maths and science curriculum in Years 3 and 5. A character called Andy poses questions to three virtual game-show contestants who demonstrate correct and incorrect thinking based on common misconceptions. Pupils answer questions as if they are taking part in a game show. Each session involves up to six minutes of science and up to six minutes of maths questions (whether maths or science comes first is determined at random). Each topic starts with an 'exploratory question'. An answer cannot be entered straight away: there is enforced stopping and thinking time during which an image of a hand pulses on the screen. After this time has elapsed, if the pupils provide a correct answer they are shown three contestants giving their thoughts about the question (one with correct reasoning, two with incorrect reasoning) and the pupils have to select the correct reasoning. When the correct reasoning is selected, the game moves on to 'practice questions' (Gauthier et al., 2022). There are four types of practice tasks for each concept which help consolidate understanding, generalise it to new stimuli, extend it to related concepts, and avoid misconceptions to prevent pupils from expecting 'trick' questions or doubting their initial thoughts. The intervention is delivered to pupils at the start of maths and/or science lessons by their teacher or teaching assistant. Figure 1 shows screenshots from the programme.

[1] Made up of experts in neuroscience, psychology and education from Birkbeck, the UCL Institute of Education and Learnus: http://unlocke.org/team.html

Figure 1: Screenshots from the Stop and Think programme



What number is shown by the blocks?

43

I have counted seven pieces in total, so that must be the answer.

There are four long bars and three spare cubes.

The four bars each have ten cubes and there are three more units to add.

The intervention was designed to be a whole-class activity, with pupils working through the problems together as a group. Teachers were able to decide how the pupils interacted with the software to input their answers, as long as the process of selecting was not based on the first pupil who responds (as this would undermine the 'stop and think' process). Teachers could also ask pupils to discuss the problems in pairs rather than as a group.

In this trial, teachers had some flexibility over what they changed in the sessions, especially around how the pupils interact with the software as a group. The change was made to empower teachers to integrate the activity within their normal class practice, thereby enhancing compliance and uptake. This flexibility was a change from the efficacy trial which had followed a more structured approach. Now, there were two optional ways that teachers could tailor their use of the software:

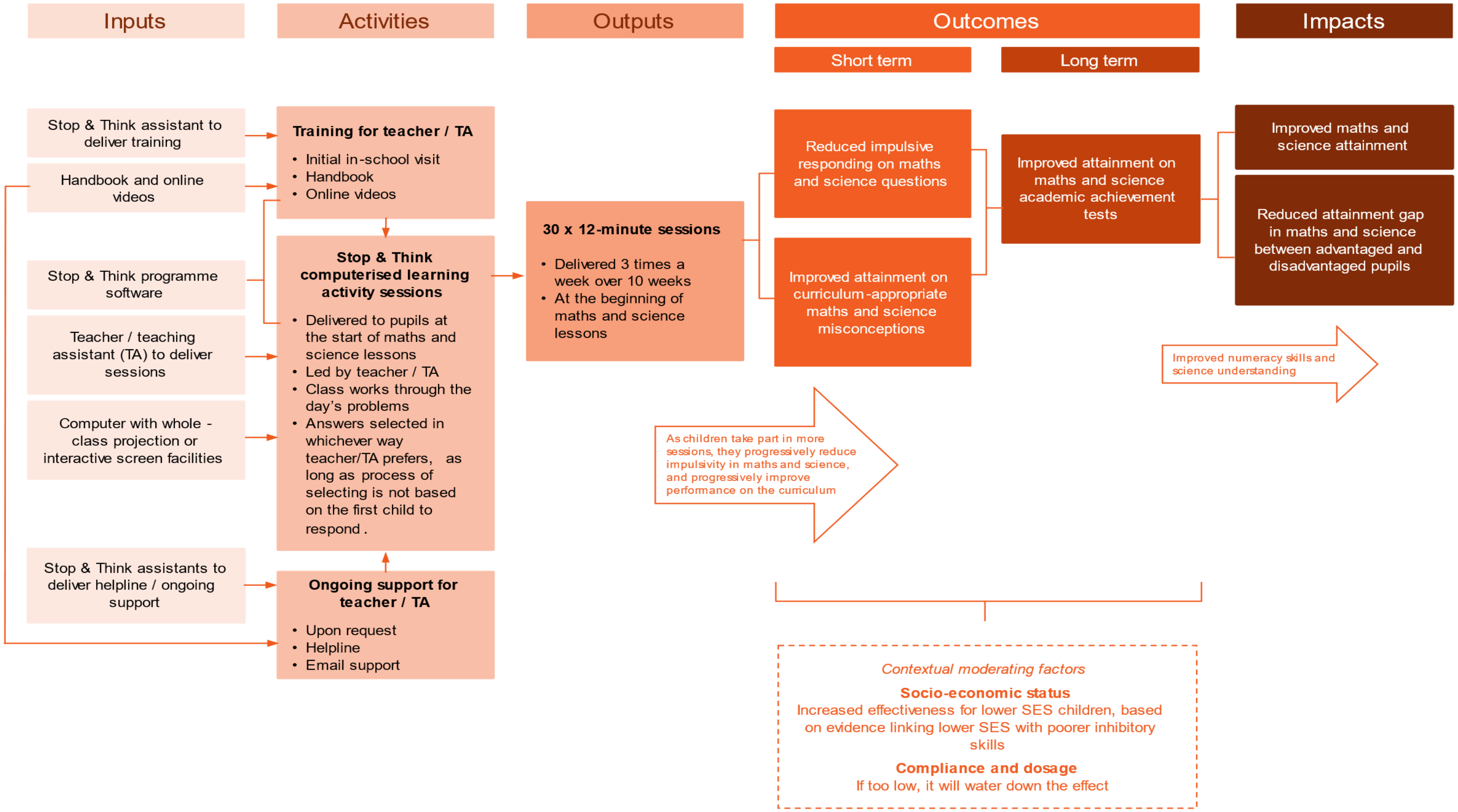
- Teachers could choose **weekly themes** (such as animals or fractions) if they wanted sessions that matched with what they wanted to teach in a specific week. Alternatively, they could opt for the random allocation of themes (which had been what the teachers did during the efficacy trial).
- Teachers could opt to include **motivational elements** in the software. One was to include the class in a leader board with other schools so pupils could see how many sessions they had completed and how they stood in relation to other schools. Pupils could only see the schools immediately above and below them rather than their overall ranking. Another motivational element was for the class to receive virtual coins for each session completed. These coins could then be used to buy items to improve an animal avatar.

Developers expected these options to encourage use of the software (that is, increase compliance and dosage) but did not expect teachers' choices about how to use the software themselves to moderate the impacts of the software on pupils' attainment in maths and science. This was because the incentivisation changes took place outside the core Stop and Think sessions and therefore would not influence impacts by affecting the core mechanism of 'stopping and thinking' as set out in the logic model (Figure 2).

Intervention logic model

Building on the logic model developed for the efficacy trial, an IDEA workshop was held in September 2020 and was attended by representatives from NatCen, Birkbeck, and BIT. Following this, we developed an updated logic model for the Stop and Think effectiveness trial in collaboration with Birkbeck (Figure 2).

Figure 2: Stop and Think logic model



Inputs

Activities

Outputs

Outcomes

Short term

Long term

Impacts

Stop & Think assistant to deliver training

Handbook and online videos

Stop & Think programme software

Teacher / teaching assistant (TA) to deliver sessions

Computer with whole - class projection or interactive screen facilities

Stop & Think assistants to deliver helpline / ongoing support

Training for teacher / TA

- Initial in-school visit
- Handbook
- Online videos

Stop & Think computerised learning activity sessions

- Delivered to pupils at the start of maths and science lessons
- Led by teacher / TA
- Class works through the day's problems
- Answers selected in whichever way teacher/TA prefers, as long as process of selecting is not based on the first child to respond .

Ongoing support for teacher / TA

- Upon request
- Helpline
- Email support

30 x 12-minute sessions

- Delivered 3 times a week over 10 weeks
- At the beginning of maths and science lessons

Reduced impulsive responding on maths and science questions

Improved attainment on curriculum-appropriate maths and science misconceptions

Improved attainment on maths and science academic achievement tests

Improved maths and science attainment

Reduced attainment gap in maths and science between advantaged and disadvantaged pupils

As children take part in more sessions, they progressively reduce impulsivity in maths and science, and progressively improve performance on the curriculum

Improved numeracy skills and science understanding

Contextual moderating factors

Socio-economic status

Increased effectiveness for lower SES children, based on evidence linking lower SES with poorer inhibitory skills

Compliance and dosage

If too low, it will water down the effect

Evaluation objectives

Our integrated evaluation design includes both an impact evaluation and an implementation and process evaluation (IPE). The evaluation protocol and statistical analysis plan are published on the **EEF's website.**[2]

The impact evaluation was designed to build on findings from the Stop and Think efficacy trial and subsequent sensitivity analysis. The evaluation was conducted 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. It aimed to answer the following 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?

It also aimed to answer the following 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?

The IPE research questions were as follows.

IPE RQ1 To what extent do the delivery partners and teachers deliver Stop and Think as intended?

IPE RQ2 How well is Stop and Think delivered?

IPE RQ3 How, why, and to what extent are changes made to Stop and Think?

IPE RQ4 Do teachers deliver the intended dose of three 12-minute sessions for ten weeks?

IPE RQ5 What is the rate and scope of participation at a school, class, and pupil level?

IPE RQ6 How well do teachers and pupils engage with the intervention?

IPE RQ7 How different is Stop and Think to usual Key Stage 2 maths and science teaching?

IPE RQ8 What outcomes do teachers and pupils perceive to result from Stop and Think?

Ethics and trial registration

NatCen has a robust ethics governance procedure. NatCen's internal Research Ethics Committee (REC) completed a full review of this trial. Ethical approval was granted in November 2020. The trial was registered on 7 May 2021 and the International Standard Randomised Controlled Trial Number (ISRCTN) is ISRCTN12838371.[3] The trial registry will be updated with outcomes at the end of the project.

Data protection

NatCen stored and handled all data securely and confidentially in line with the U.K. General Data Protection Regulation (UK GDPR). Only the research team had access to data collected as part of the evaluation. This was monitored through a data security plan, which detailed all data security procedures to be applied, including names of those who have access rights to respondent confidential data, details of third parties (such as transcribers) involved in the project and specific requirements for data destruction. NatCen issued a privacy notice to all concerned parties, which was also

[2] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/stop-and-think-learning-counterintuitive-concepts-regrant

[3] **ISRCTN registry,** https://doi.org/10.1186/ISRCTN12838371**.**

published on the study website. School and pupil-level data was transferred to NatCen via a secure file transfer service. The privacy notice stated that reports and other publications which arise from this research will not identify any individual school, staff member, or pupil. It also mentioned that schools or individual staff who no longer wished to take part in the evaluation could request to have their data deleted at any point prior to the submission of the draft report.

During the study, NatCen was the data controller and additionally processed data. The legal basis for processing the data following GDPR Article 6 was 'legitimate interest'. NatCen processed the data for the legitimate purpose of conducting the evaluation of the Stop and Think programme. No special category data was collected as part of the evaluation. After the evaluation report is published, data from the impact evaluation will be stored in the EEF archive. At this point, the EEF becomes the data controller. Personal identifiers were removed and replaced by a meaningless identifier (so-called Pupil Matching Reference, 'PMR') and uploaded to the Secure Research Service (SRS) hosted by the Office for National Statistics (ONS). The resulting dataset cannot be used to identify any individual pupil in the SRS.

All evaluation data will be securely deleted from NatCen's systems 12 months after the end of the trial (by September 2025 at the latest).

Project team

The evaluation was carried out by education and evaluation specialists at NatCen who brought together extensive experience of school-based research, including of other large-scale trials for the EEF.

Table 1: Evaluation team

Name	Project role	Role and team
Helena Takala	Principal investigator and project manager (from August 2023)	Research Director, Centre for Children and Families
Mary McKaskill	Principal investigator and project manager (until August 2023)	Research Director, Centre for Children and Families
Enes Duysak	Impact evaluation lead	Research Director, Centre for Evaluation
Tien-Li Kuo	Impact evaluation support	Senior Researcher, Centre for Evaluation
Ekaterina Stoilova	Impact evaluation support	Senior Researcher, Centre for Evaluation
Sehaj Bhatti	IPE and endline testing lead	Senior Researcher, Centre for Children and Families
Nicky McGuinness	IPE support	Senior Researcher, Centre for Children and Families
Alina Fletcher	IPE support	Senior Researcher, Centre for Children and Families
Daniel Finn	Impact evaluation data management	Data Director, Data Management
Andi Fugard	Impact evaluation QA	Co-Director, Centre for Evaluation
Gayle Munro	IPE QA	Director, Centre for Children and Families

Table 2: Delivery team

Name	Project role	Role and team
Denis Mareschal	Developer and academic advisor	Professor of Psychology, Centre for Brain and Cognitive Development, Birkbeck University of London
Kathryn Atherton	Delivery lead	Senior Advisor, Education, BIT
Callum O'Mahony	Delivery support	Advisor, Education, BIT
Julia Ryle-Hodges	Delivery support	Advisor, Education, BIT
Hannah Bellier	Project Champion (Stop and Think research assistant)	Associate Advisor, BIT
Martha Courtauld	Project Champion (Stop and Think research assistant)	Research Assistant, BIT
Priya Chahal	Project Champion (Stop and Think research assistant)	Research Assistant, BIT
Paige Lindsey	Project Champion (Stop and Think research assistant)	Research Assistant, BIT
Fionnuala O'Reilly	Delivery lead (until September 2022)	Senior Advisor, Education, BIT
Dave Wilson	Delivery support (until September 2022)	Advisor, Education, BIT
Anna Bird	Project oversight	Principal Advisor, Education, BIT
Lal Chadeesingh	QA	Principal Advisor, Education, BIT

Methods

Trial design

The evaluation was designed as a two-arm cluster randomised controlled effectiveness trial of the impact of Stop and Think on Year 3 and Year 5 maths and science attainment. As an effectiveness trial, the evaluation aimed to test the effect of the intervention in 'real-world' circumstances. Table 3 summarises the design of the cluster-RCT.

Schools taking part in the trial were randomly assigned into two conditions. In both conditions, one year group (Year 3 or Year 5) received the intervention while the other year group continued teaching as usual. For schools randomised into Condition 1, all Year 3 pupils, across classes, received the Stop and Think programme in maths and/or science lessons while all Year 5 pupils continued with teaching as usual. For schools randomised into Condition 2, all Year 5 pupils, across classes, received the Stop and Think programme in maths and/or science lessons while all Year 3 pupils continued with teaching as usual. This design ensured that each school received the intervention, reducing the number of schools needed for the trial making initial recruitment easier and maintaining low school-level attrition.

The randomisation was stratified by class-form entry size (one, two, or three or more classes per year group) and the school-level proportion of FSM-eligible pupils (by tercile of distribution) to ensure balance on these school characteristics across conditions.

Additionally, pupils in all year groups and classes were individually randomly assigned to sit either maths or science tests. This meant that 50% of pupils in each year group and each class were tested in maths attainment and maths misconceptions and 50% in science attainment and science misconceptions. This approach reduced the testing burden on pupils and provided greater statistical power than school-level randomisation to either maths or science testing. Table 3 illustrates this two-stage randomisation approach; Table 4 presents the overall trial design.

Table 3: Trial design—condition and test allocation

School allocation	Year group	Trial arm	Baseline measures	Outcome tests*
Condition 1	Year 3	Stop and Think (maths + science)	KS1 maths attainment	PTM8
				PTS8
	Year 5	Control: teaching as usual	EYFSP overall point score	PTM10
				PTS10
Condition 2	Year 3	Control: teaching as usual	KS1 maths attainment	PTM8
				PTS8
	Year 5	Stop and Think (maths + science)	EYFSP overall point score	PTM10
				PTS10

* Age-specific GL Progress Tests in Maths (PTM) and Science (PTS) were used for outcome tests; for each test, 50% of the pupils were tested.

The outcome measures selected for the trial were selected on the basis of their age-appropriateness and match with the programme logic model. They were selected in collaboration with the delivery team. The primary outcome of interest for this evaluation was maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds, defined as those who have been eligible for FSM at any point in the previous six years. The trial focused on this primary outcome because addressing pupil disadvantage is a key priority area for the EEF.

The secondary outcomes in this trial were 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. To measure maths and science attainment following intervention delivery we used the age-specific GL Progress Tests in Maths and Science (PTM8 and PTS8 for pupils in Year 3; PTM10 and PTS10 for pupils in Year 5). To measure the prevalence of common misconceptions in maths and science we used age- and subject-specific tests developed by NatCen and Oxford MeasurEd. More details on the measures are provided in the Outcome Measures section.

Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, identical baseline measures for Year 3 and Year 5 pupils were not available. The trial therefore used 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 the baseline measures are provided in the Outcome Measures sub-section.

Table 4: Trial design

Trial design, including number of arms		Two-arm, cluster randomised controlled trial.
Unit of randomisation		School level for condition allocation.
Stratification variable(s) (if applicable)		Class-form entry (whether there are 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 among FSM pupils.
	Measure (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), GL Assessment; Year 5: Progress Test in Maths (PTM10), GL Assessment.
Secondary outcome(s)	**Variable(s)**	Maths attainment among all pupils; science attainment among FSM pupils and among all pupils; common misconceptions in maths and science among all pupils.

	Measure(s) (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), GL Assessment;
		Year 5: Progress Test in Maths (PTM10), GL Assessment;
		Year 3: Progress Test in Science (PTS8), GL Assessment;
		Year 5: Progress Test in Science (PTS10), GL Assessment;
		both years: age-specific common misconceptions in maths and science tests (developed by Oxford MeasurEd and NatCen).

Baseline for primary outcome	**Variable**	Year 3: maths attainment; Year 5: EYFSP overall progress*
	Measure (instrument, scale, source)	Year 3: KS1 maths attainment, eight-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(s)	**Variable**	Year 3: maths attainment; Year 5: EYFSP overall progress.
	Measure (instrument, scale, source)	Year 3: KS1 maths attainment, eight-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.

* Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, the trial used 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.

Participant selection

School recruitment

BIT identified and recruited eligible schools between October 2021 and July 2022. All state primary schools in England were eligible for the trial unless (a) pupils from Year 3 and Year 5 were taught in the same class, (b) the school had previously partnered with the Stop and Think programme, or (c) the school had taken part in piloting work for the misconceptions tests that were used as a secondary outcome measure. Schools were asked to sign a Memorandum of Understanding (MoU) confirming their commitment to delivering the programme as required and taking part in evaluation activities. This process was completed before the start of the 2022/2023 academic year. Information sheets shared with schools and parents are available in the trial protocol.

Pupil recruitment

All Year 3 and Year 5 pupils in recruited schools were eligible for the trial. Participating schools shared the trial information leaflet and privacy notice with all Year 3 and Year 5 pupils and their parents/carers. Parents/carers were given two weeks to withdraw from the trial and from data processing after which schools were asked to share with NatCen pupil-level information for all Year 3 and Year 5 pupils who had not been formally withdrawn from the trial. These pupils made up the trial participants.

Schools provided background information for all trial participants, including Unique Pupil Number (UPN), date of birth, first name, surname, and class in multi-form schools for all Year 3 and Year 5 pupils. This pupil information was collected in an Excel spreadsheet template and uploaded by schools using a secure portal on the NatCen website.

Sample size

We designed this trial to be powered to detect an effect of the size found in post-hoc analysis of the efficacy trial[4] on the primary outcome (maths attainment among Year 3 and Year 5 pupils eligible for FSM) and on the secondary outcomes (science attainment and maths attainment for all Year 3 and Year 5 pupils).

At the protocol stage, after accounting for expected attrition at both school and pupil level, we assumed a high pupil-level correlation between baseline and endline testing (0.635) and a small school-level intra-cluster correlation (ICC = 0.07 for the primary outcome analysis). Our power calculations were informed by the post-hoc analysis of the efficacy trial.[5] We also made a conservative assumption that a school-level correlation between baseline and endline is 0.0 and used a Type I error rate of 0.05 and a Type II error rate of 0.20 (power of 0.80). The planned recruitment of 165 schools (n = 1,156 pupils) would yield an MDES of 0.17 for the primary analysis of maths attainment among KS2 pupils eligible for FSM. Details are covered in the trial protocol under the Power Calculation section.[6] We conducted this and the following power calculations using the *PowerUp!* tool (Dong and Maynard, 2013).

At the randomisation stage, we used the number of schools retained in the study at that point (n = 173). We calculated an average cluster size of 42.5 pupils per year group per school. The calculations used the actual data at that time and were based on the same core assumptions as the protocol stage.

At the analysis stage, we updated the power calculations using the final number of pupils included in the endline primary analysis (n = 1,841) to allow comparison between the number of pupils included in the study at randomisation and protocol stages (n = 2,249; n = 1,156 respectively). We also used the updated Level 1 pre-test/post-test correlations (r = 0.51 for the primary analysis compared to 0.635 at the protocol stage; the R^2 value was used for input into PowerUp!) and Level 2 pre-test/post-test correlations (r = 0.096 for the primary analysis compared to 0.0 at the protocol stage) for the power calculations. Additionally, we updated the ICCs based on the endline analysis. With these updated assumptions, the overall MDES at endline is 0.15 standard deviations for the primary analysis of maths attainment among KS2 pupils eligible for FSM. Table 5 and Table 6 present our sample size calculations by maths and science for this trial from the protocol stage to randomisation of condition allocation, and to analysis.

[4] Durham University (2020) ‘Re-analysis: Stop and Think: Learning Counterintuitive Concepts’. Unpublished manuscript.

[5] 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.

[6] ‘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

Table 5: Minimum detectable effect size for maths

		Protocol		**Randomisation**		**Analysis**	
		All pupils	**FSM (primary analysis)**	**All pupils**	**FSM (primary analysis)**	**All pupils**	**FSM (primary analysis)**
MDES		0.13	0.17	0.12	0.14	0.14	0.15
Pre-test/post-test correlations	Level 1 (pupil)	0.635	0.635	0.635	0.635	0.55	0.51
	Level 2 (school)	0.0	0.0	0.0	0.0	0.187	0.096
Intracluster correlations (ICCs)	Level 2 (school)	0.07	0.07	0.07	0.07	0.09	0.05
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size		15.7	3.5	21.3	6.5	18.1	5.6
Average year group size7		31.4	7	42.5	13	36.28	11.2
Number of schools	Treatment	165	165	173	173	167	164
	Control	165	165	173	173	167	164
	Total	165	165	173	173	167	1649
Number of pupils	Treatment	2587	578	3676	1125	3080	943
	Control	2587	578	3676	1125	2964	898
	Total	5174	1156	7353	2249	6044	1841

[76] '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

[8] The average year group size for analysis is derived from the total number of pupils divided by the total number of schools while the harmonic mean of number of pupils per school is 43 for all maths and ten for FSM maths.

[9] 164 of the 167 schools for maths analysis had at least one pupil who was eligible for FSM and whose baseline score and PTS were not missing.

Table 6: Minimum detectable effect size for science

		Protocol		Randomisation		Analysis	
		All pupils	FSM	All pupils	FSM	All pupils	FSM
MDES		0.14	0.18	0.14	0.16	0.14	0.16
Pre-test/post-test correlations	Level 1 (pupil)	0.645	0.645	0.645	0.645	0.52	0.47
	Level 2 (school)	0.0	0.0	0.0	0.0	0.199	0.066
Intracluster correlations (ICCs)	Level 2 (school)	0.09	0.09	0.09	0.09	0.09	0.06
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size		15.7	3.5	21.3	6.5	18.2	5.7
Average year group size[10]		31.4	7	42.5	13	36.3[11]	11.4
Number of schools	Treatment	165	165	173	173	166	159
	Control	165	165	173	173	166	159
	Total	165	165	173	173	166	159[12]
Number of pupils	Treatment	2587	578	3676	1125	3062	923
	Control	2587	578	3676	1125	2971	896
	Total	5174	1156	7353	2249	6033	1819

Baseline and outcome measures

Baseline measures

Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, identical baseline measures for Year 3 and Year 5 pupils were not available. We used KS1 maths attainment as a baseline measure of maths attainment

[10] 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.

[11] The average year-group size for analysis is derived from total number of pupils divided by total number of schools while the harmonic mean of number of pupils per school is 43 for all science and 11 for FSM science.

[12] 159 of the 166 schools for science analysis had at least one pupil who was eligible for FSM and whose baseline score and PTS were not missing.

for Year 3 pupils. As national testing in schools was cancelled due to the COVID-19 pandemic in 2019/2020 when this cohort of pupils were five years old, no EYFSP data is available for this cohort.

The cancellation of national testing in 2019/2020 also prevented us from using KS1 test scores as prior attainment measures for the Year 5 cohort. We therefore used the EYFSP overall point score[13] as the baseline measure of attainment for Year 5 pupils. As the baseline measures for Year 3 and Year 5 pupils had different scales, we standardised the baseline measure to have a mean of zero and standard deviation of one by each year group. We then combined them to form a single baseline measure.

We obtained these baseline measures along with FSM eligibility information for all trial participants from the National Pupil Database (NPD) after receiving the pupil sample in autumn term 2022.[14]

Primary outcome

The primary outcome for this evaluation was maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds, defined as those who had been eligible for FSM at any point in the previous six years. The trial focused on this primary outcome for multiple reasons. First, addressing pupil disadvantage is a key priority for the EEF; second, post-hoc analysis of the Stop and Think efficacy trial[15] found a significant and comparatively large effect of the programme among FSM-eligible pupils across year groups 3 and 5; last, the efficacy trial found a positive but not significant effect on maths attainment.

No relevant national tests were available for Year 3 and Year 5 pupils through the NPD as age-specific tests for maths and science attainment. As a primary outcome measure we therefore used the GL Assessment Progress Test in Maths (GL PTM)[16] age-standardised scores as a standardised measure of pupils' mathematical skills and knowledge.[17] Age-appropriate versions of the paper-based test were administered to each year group (PTM8 for Year 3; PTM10 for Year 5 pupils) between May and July 2023. PTM8 and PTM10 tests consist of two parts: (a) Mental Maths and (b) Applying and Understanding Maths. Mental Maths questions were timed and played from an audio file while Applying and Understanding Maths questions were answered at the student's own pace.

Independent test administrators from NatCen visited schools to conduct endline testing. They were recruited from NatCen's field interviewer panel and briefed by the research team on how to administer and invigilate the testing. To reduce the burden on pupils, we randomised pupils in each school and year group/class to sit either maths or science outcome tests (see Randomisation section). Pupils sat the tests in their classroom under test-like conditions in May to July 2023. Each individual pupil either took science or maths tests: one GL Progress Test in science or maths and the corresponding misconception test (that is, pupils who took the maths GL Assessment also took the maths misconception test, and those who took the science GL test also took the science misconception test: see Secondary Outcomes section). Pupils took a break between the two tests. Invigilators scheduled two-hour testing sessions per classroom, which included set-up, administration, and break times. The GL Assessment Progress Test took 45 minutes and the misconception test 35 minutes. The GL Progress Test in Maths included an additional audio recording section where a 7- to 8-minute audio was played for the pupils and they were asked to answer questions based on that. This section of the test was adjusted and delivered at the end of the misconceptions test in order to avoid disturbances to the classroom. Invigilators were asked not to break before the full time was up, even if it appeared that everyone was finished, in order not to rush pupils through the questions; this would have contravened the ethos of the programme to 'stop and think'.

[13] Following the efficacy trial, we combined all 17 early learning goals to obtain an average EYFSP point score. Early learning goal variables were obtained from the NPD and take a value between one and three, where higher scores reflect higher attainment for each specific learning goal.

[14] The Office of Qualifications and Examinations Regulation, Department for Education (2022), ‘GRading and Admissions Data England-Ofqual-DfE’, available at https://doi.org/10.57906/4phz-dq28

[15] Durham University (2020) ‘Re-analysis: Stop and Think: Learning Counterintuitive Concepts (137)’, unpublished.

[16] For more information, please see GL Assessment (no date) ‘Progress Test in Maths’, available at https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/

[17] Due to the high number of missing information for pupil’s age, one pseudo-birth date was assumed for all pupils in each year group and scores were standardised using this date (instead of actual birthdays).

The test invigilators collected completed test papers and returned them back to NatCen on completing school visits. The GL Progress Tests were transcribed and scored by GL Assessment while the misconception tests were transcribed by a company called Adetiq.

Secondary outcomes

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 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 for all pupils

As secondary outcome measures, we used the GL Assessment Progress Test in Maths (GL PTM) to evaluate maths attainment for all Year 3 and Year 5 pupils as well as the GL Assessment Progress Test in Science (GL PTS) as a standardised measure of pupils' science skills and knowledge. GL Assessment PTS age-standardised scores were used to measure science attainment for all Year 3 and Year 5 pupils and among only pupils eligible for FSM only.[18] Age-appropriate versions of this test were administered to each year group (PTS8 for Year 3; PTS10 for Year 5) alongside maths assessments (with 50% of pupils randomly allocated to take maths and 50% science assessments, respectively, see Randomisation section).

Common misconceptions in maths and science

To assess the effect of Stop and Think on the prevalence of common misconceptions—one of the short-term outcomes in the logic model—NatCen and Oxford MeasurEd developed bespoke tests for common misconceptions in maths and science. These were used as secondary outcome measures.

Four tests were developed (one test per subject per year group, that is, Year 3 maths, Year 3 science, Year 5 maths, and Year 5 science) testing pupils' tendency to fall into common misconceptions in maths and science. The tests consist of multiple-choice items. For each item, one of the incorrect responses includes a common misconception. Details on test development and validation are available in a technical report to be published alongside the present evaluation report (McKaskill et al., 2025). The technical report presents the rationale, methodology, and process for item development, validation, as well as item inclusion or exclusion from analysis for the Stop and Think effectiveness trial.

During endline testing, pupils randomly allocated to take the GL Assessment Progress Test in maths took the age-appropriate test for common misconceptions in that subject, while pupils randomly allocated to take the GL Assessment Progress Test in science took the science equivalent. The final version of common misconception tests that was used during endline testing consisted of 15 items for Year 3 maths, Year 5 maths, and Year 5 science and 16 items for Year 3 science.

At the analysis stage, five items were validated and included in the secondary analysis for Year 3 maths and nine for Year 5 maths. Seven items were validated and included in the secondary analysis for Year 3 science and nine for Year 5 science. Following the validity and reliability assessment, (McKaskill et al., 2025) identified items reliably capturing misconceptions. We only used responses to these validated items for the analysis of this secondary outcome measure.

For the analysis, we defined the common misconceptions in maths and science as the number of times the pupil fell into a common misconception out of the total number of validated items per year group. In our analyses, we used the raw scores depicting this value for maths and science, separately.

[18] Due to the high number of missing information for pupils' ages, one pseudo-birth date was assumed for all pupils in each year group and scores were standardised using this date (instead of actual birthdays).

Randomisation

School-level condition allocation

Randomisation for treatment allocation in the trial was at the school level such that schools were randomised into one of two conditions that determined which year group received the programme and which did not. For schools in Condition 1, Year 3 pupils received the Stop and Think programme while Year 5 pupils continued teaching as usual, and for schools in Condition 2, Year 5 pupils received Stop and Think while Year 3 pupils continued teaching as usual (see Table 3). Hence, in both conditions, each school had one Stop and Think year group and one teaching-as-usual year group. This way of randomising ensured that every school received the programme, making recruitment easier and reducing the likelihood of school drop-outs. Additionally, it provided greater statistical power than whole-school randomisation to either intervention or control. As described above (see Sample Size), a total of 173 schools were included for randomisation (n = 14,718 pupils).

Randomisation at school level was stratified by class-form entry size (whether there is one, two, or three or more classes per year group) and the school-level proportion of FSM-eligible pupils (by tercile of distribution) to ensure balance on these school characteristics across conditions.

Randomisation of schools was carried out by the impact evaluation team at NatCen, with the researcher blinded to school identity, using the `randtreat` command with the `misfits(global)`option in Stata version 17 on 13 October 2022.

This process assigned 87 schools to Condition 1 (with Year 3 pupils receiving Stop and Think) and 86 schools to Condition 2 (with Year 5 pupils receiving Stop and Think; see Participant flow). We communicated the outcome of school randomisation to the delivery team which, in turn, notified schools. Table 7 further presents the number of classes by year group randomised to each condition and trial arm.

Table 7: Number of classes by year group, condition, and trial arm at school-level condition allocation

School condition	Year 3	Year 5	Total classes
Condition 1	152 (intervention)	148 (control)	300
Condition 2	146 (control)	147 (intervention)	293
Total: 173 schools, 593 classes (299 intervention; 294 control)			

After endline data collection, we were notified that the delivery team recorded and notified two schools of the opposite condition allocation. While this means that the number of schools by condition and that of classes by trial arm remained the same (87 schools in Condition 1 and 86 schools in Condition 2; 299 classes in intervention and 294 classes in control). This is a two-sided non-compliance, which was explored in the compliance analysis (see Compliance).

Pupil-level test assignment

Pupils in each school, year group, and class were randomised to sit either maths or science outcome tests: 50% of pupils in each school, year group and class were tested in maths attainment and maths misconceptions, and 50% in science attainment and science misconceptions (see Table 3).

Randomisation of pupil-level test assignment was stratified by school, year group, and class, and was carried out by the impact evaluation team at NatCen, with the researcher blind to pupil identity, using the `randtreat` command in Stata version 17 on 8 March 2023. The procedure followed was analogous to the one described above, with the exception that within each stratum, schools were listed in descending order by NatCen-created unique school identifiers and by year group.

Randomisation of pupil-level test assignment was carried out with 170 schools and 14,645 pupils (see Attrition). More details on the rationale behind our randomisation approach are provided in the Stop and Think protocol[19] and more details on randomisation (for example, numbers by characteristics) can be found in the trial's Statistical Analysis Plan.[20]

Statistical analysis

The outcome analysis was undertaken on an intention-to-treat (ITT) approach, with schools analysed as per intended condition allocation to Condition 1 (with Year 3 pupils receiving Stop and Think) or Condition 2 (with Year 5 pupils receiving Stop and Think).

Primary analysis

The primary outcome analysis addressed our 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?

A two-level model was used to account for pupils (level one) being clustered in schools (level two) and included year groups as a fixed effect to estimate a pooled effect of the programme for both year groups.[21]

Age-standardised PTM score from age-appropriate maths tests was the dependent variable in this model, with a binary treatment allocation indicator, standardised baseline scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils), year group, and randomisation strata as covariates. School-level random effects were included in the model by allowing the intercept to vary by school. 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 baseline and endline attainment scores; and

β_2 is the average effect of the programme on PTM scores.

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.

School-level random effects account for school-level variation in the outcome not explained by the fixed effects. The analysis was implemented in Stata 17 using the `mixed` command.

Following the SAP and EEF statistical analysis guidance (EEF, 2022), we used a restricted sample of pupils eligible for FSM for the primary analysis. For the robustness check, we employed an interaction model incorporating the FSM eligibility indicator and an interaction term that combined FSM eligibility with treatment allocation in the further sensitivity analysis. Details on the interaction model are provided in the Further Sensitivity Analysis section.

The impact of the programme is expressed as a standardised effect size. See the Effect Size calculation section below for an explanation of how effect sizes were calculated.

[19] Stop and Think: Learning Counterintuitive Concepts Evaluation Protocol, retrieved from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Stop-Think_trial_protocol_280322_v1.pdf?v=1714392919

[20] Stop and Think: Learning Counterintuitive Concepts Statistical Analysis Plan, retrieved from: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/EEF-SAP-Stop-and-Think_v1.pdf?v=1714392919

[21] As class-level clustering/grouping is more common in secondary schools than in primary schools in England (Dracup, 2014), class-level were not considered during the design stage.

Secondary analysis

For all secondary outcome analyses, models included a binary indicator of treatment allocation and the randomisation strata and year group as fixed effects. The analyses were implemented in Stata 17 using the `mixed` command.

Maths and science attainment

The secondary outcome analysis for maths and science attainment addressed the following 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?

Using the same approach as outlined for the primary outcome analysis above, a two-level model was estimated for each secondary outcome for both Year 3 and Year 5 pupils combined to reflect pupils (level 1) nested within schools (level 2). Age-standardised PTM or PTS scores were the dependent variable and standardised baseline attainment scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) were included as a covariate. 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 baseline and endline attainment scores; and

β_2 is the average effect of the programme.

The term u_j is a school-level random effect and e_{ij} is the error term, both assumed to be normally distributed and uncorrelated with all the covariates included in the model.

The impact of the programme is expressed as a standardised effect size. See the Effect size calculation section for an explanation of how effect sizes were calculated.

Maths and science misconceptions

The secondary outcome analysis for maths and science misconceptions addressed the following research questions:

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?

Following suggestions from peer reviewers at the SAP stage, we decided to use two-level Poisson regression models to analyse misconceptions as count measures. However, we applied the same approach as used for the primary and secondary outcome analyses for the misconception analysis. This deviation was to acknowledge that misconception measures are scales rather than counts, and that each instance of misconception was not empirically independent, which is a key assumption of Poisson models.

A two-level model was estimated for each misconception outcome for both Year 3 and Year 5 pupils combined. The misconception scores (measured as the number of times the pupils fell into a misconception) were the dependent variable, and baseline attainment scores were included as covariates. The basic form of the model for both year groups combined is as follows:[22]

$$Outcomes_{Misconception_{ij}} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + \beta_3 Stratification'_{ij} + \beta_4 YearGroup_{ij} + u_j + e_{ij}$$

where

pupils (*i*) are clustered within schools (*j*);

β_0 is an overall intercept; and

β_2 is the average effect of the programme.

The term u_j is a school-level random effect and e_{ij} is the error term, both assumed to be normally distributed and uncorrelated with all the covariates included in the model.

The impact of the programme on misconceptions is expressed as a standardised effect size. See the Effect Size Calculation section for an explanation of how effect sizes were calculated.

Additional analyses

Sensitivity analysis for the primary outcome

Similarly to the primary outcome analysis, our sensitivity analysis followed the approach adopted in the efficacy trial (Roy et al., 2019).

For maths attainment as the outcome, two-level fixed effects models were estimated for each year group separately to reflect pupils (level 1) nested within schools (level 2). Each model included raw PTM scores as the dependent variable, and standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate, as above. The basic form of the model for each year group is:

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

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 report confidence intervals at 95% level and the effect size using Hedges' g formula as described below.

To calculate a single effect size for both Year 3 and Year 5 pupils that is comparable with findings from other studies, including the Stop and Think efficacy trial, we took the mean of the two effect sizes from the two separate models for each year group. The variance of the combined effect size was estimated using the formula in Borenstein, Hedges, Higgins and Rothstein (2009, p. 218), following the EEF statistical analysis guidance.

[22] There are five validated items for the Year 3 maths misconception test and nine for the Year 5 maths while there are seven items for Year 3 science and nine for Year 5 science. Since we have already included year group in the model, this would adjust for the amount of opportunity a misconception event had. Hence, we did not include an offset term in the Poisson regression model to account that pupils' exposure to risk of falling into misconceptions which vary by test (that is, year group or subject).

To calculate a precise estimate of the overall effect size for both Year 3 and Year 5, we assigned 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). With 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

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

We initially intended to conduct the sensitivity analysis solely for the primary outcome (maths pupils eligible for FSM). Nevertheless, we extended the sensitivity analysis to include all Year 3 and Year 5 maths pupils (RQ2), all Year 3 and Year 5 science pupils (RQ3), and science pupils eligible for FSM (RQ4). This approach allows for a comparison of findings with the efficacy trial and helps contextualise the teacher comments as reported in the IPE.

Further sensitivity analysis

To assess whether findings for the primary analysis are robust to different model specifications, we conducted further sensitivity analysis. An alternative two-level model was estimated for the impact of the programme on the sample as a whole, including an interaction term between treatment status and a dummy variable indicating FSM eligibility status.

The model included standardised pre-trial test scores (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:

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

where

pupils eligible for FSM (*i*) are clustered within schools (*j*);

β_4 is the attainment gap—the difference in the average effect of the programme between FSM pupils and their peers;

β_2 is the impact of the programme on non-FSM pupils; and

the impact of the programme 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 were included as fixed effects in this model, while the school-level random effects controlled for other observed and unobserved school-level characteristics. The analysis was implemented in Stata 17 using the `mixed` command.

Additionally, we estimated the effect size associated with the effect of the programme on the attainment gap (that is, the difference in PTM scores between FSM and non-FSM pupils). Following EEF statistical analysis guidance (EEF, 2022), we reported the interaction term coefficient (β_4) and its associated measure of uncertainty and estimated the effect size based on the unconditional standard deviation of the FSM subsample, as described below (see Effect Size Calculations).

For comparability between maths and science attainment, we extended the interaction term model to include science pupils eligible for FSM. The model specification is analogous, where PTS_{ij} is the outcome variable.

Mediation analysis

Mediation analysis can help explore mechanisms by which an intervention might affect the outcomes of interest. One of the mechanisms proposed in the programme logic model by which Stop and Think could affect maths attainment among Year 3 and Year 5 pupils eligible for FSM was by reducing curriculum-appropriate maths misconceptions. Maths misconception tests were administered at endline (see Common Misconceptions in Maths and Science) and the number of items for which each pupil fell into a misconception is used for this analysis.

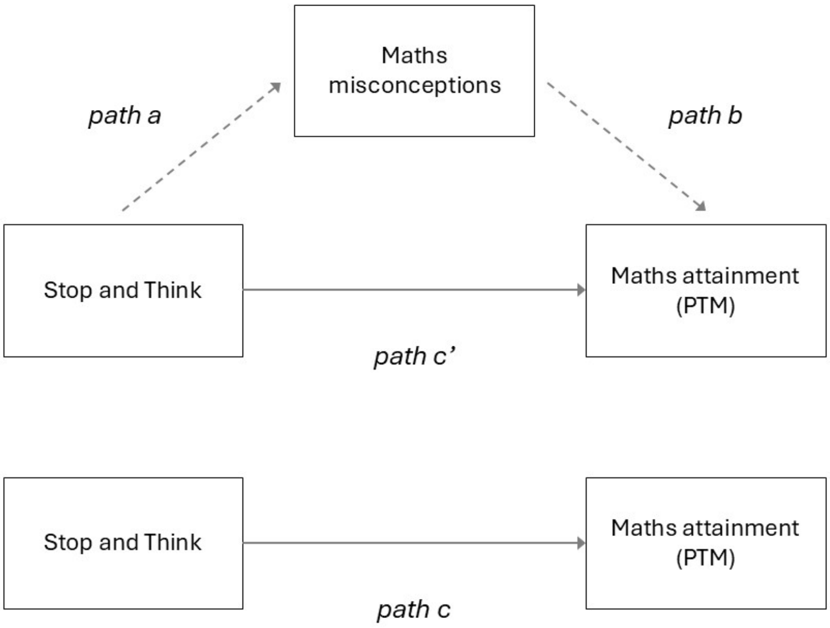
This exploratory analysis sought to decompose the intention-to-treat estimate into an indirect effect of the intervention on maths attainment (that can be attributed to changes in common misconceptions in maths) and a direct effect of the intervention on maths attainment (that cannot be attributed to changes in common misconceptions in maths). It was anticipated that a hypothesised effect of the programme on maths attainment would at least partly be mediated.

The mediation analysis followed the causal steps approach (Baron and Kenny, 1986), involving the following steps:

1. Regressing pupils' common misconceptions in maths on the Stop and Think programme. The effect of the programme on the mediator is conventionally referred to as *path a*.
2. Regressing pupils' attainment in maths on the Stop and Think programme and on pupils' misconceptions in maths. The effect of the mediator on the outcome is conventionally referred to as *path b*, while the unique direct effect of the programme (with the mediator accounted for) is referred to as *path c*'. The total effect of the programme on the outcome (direct and via mediator) is referred to as *path c*.
3. Estimating the average causal mediated effect (ACME, *path a*b*) and the proportion mediated effect (that is, the magnitude of the mediated effect relative to the total effect).

Figure 3 displays the causal mediation model explored.

Figure 3: Causal mediation model



Note: *Path a* is the effect of programme on the mediator. *Path b* is the effect of the mediator on the outcome. *Path c'* is the unique direct effect of the programme on the outcome, with the mediator accounted for. *Path c* is the total effect of the programme on the outcome, both direct and via the mediator.

To estimate *path a* of the causal mediation analysis, we estimated a two-level linear mixed effects regression model, predicting maths misconception scores—the number of times the learner fell into a misconception—from programme allocation, with year group,[23] baseline maths attainment,[24] and randomisation strata as covariates. The model included a school-level random intercept.

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

The slope β_1 represents the difference in maths misconceptions between pupils who received Stop and Think and those who did not, that is, the effect of the programme on maths misconceptions as the mediator.

As step two of the causal mediation analysis, we estimated a two-level linear mixed effects regression model predicting PTM scores from programme allocation and maths misconception scores, with year group, baseline maths attainment,[25] and randomisation strata included as covariates. The model included a school-level random intercept.

$$PTM_{ij} = \alpha_0 + \alpha_1 Intervention_j + \alpha_2 YearGroup_{ij} + \ \alpha_3 Baseline_{ij} + \alpha_4 Stratification'_j + \alpha_5 Maths\ Misconceptions_{ij} + u'_j + e'_{ij}$$

The slope α_1 provides the average direct effect (ADE) of Stop and Think on maths attainment, while α_5 represents the change in maths attainment for a unit increase in maths misconceptions, that is, the effect of the mediator on maths attainment. Drawing on the two models, $\beta_1\alpha_5$ provides the average causal mediated effect (ACME).

The ACME, ADE, and the proportion mediated effect were estimated using the `mediation` package in R (Imai et al., 2010).

$$Proportion\ mediated = \frac{ACME}{Total\ effect} = \frac{\beta_1\alpha_5}{\beta_1\alpha_5 + \alpha_1}$$

For all steps, we present unstandardised model coefficients, p values, and 95% confidence intervals obtained using quasi-Bayesian estimation with 1,000 simulations. The primary effect size interpreted is the proportion mediated effect and its confidence interval.

Imbalance at baseline

To check for, and monitor, imbalance at baseline (that is, after obtaining baseline data) we undertook descriptive analysis at school and pupil levels. Imbalance was assessed first at the school level by condition allocation, covering class-form entry size and school-level proportion of pupils eligible for FSM at any time during the previous six academic years, as school-level characteristics. Comparisons between treatment and control groups at pupil level covered FSM eligibility in the previous six years and year-group standardised baseline attainment scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils).

Categorical variables were explored by conducting cross-tabulations, including counts and percentages in each category. Continuous variables were summarised with descriptive statistics (n, mean, standard deviation, and effect size) by condition or group allocation. Standardised mean differences in baseline characteristics were calculated as Hedges' g effect sizes. An effect size greater than 0.05 was considered as an indication of possible imbalance, for which a sensitivity analysis would be estimated.

[23] There are five validated items for the Year 3 maths misconception test and nine for the Year 5 maths. Since we have already included year group in the model, this would adjust for the amount of opportunity a misconception event had. Hence, we did not include an offset term in the Poisson regression model to account that pupils' exposure to risk of falling into misconceptions which vary by test (i.e., year group or subject). This apples to the two-level linear mixed effects regression model (step two).

[24] Baseline maths attainment was included in the mediation model predicting maths misconceptions to improve precision and replicate the model used when analysing misconceptions as a secondary outcome. This is a deviation from the SAP, which did not include baseline attainment in mediation models.

[25] Baseline maths attainment was included in the mediation model predicting maths attainment (PTM) to improve precision and replicate the structure of the primary outcome model (apart from the mediator also being included as a predictor in this mediation model). This is a deviation from the SAP, which did not include baseline attainment in mediation models.

No imbalances greater than a Hedge's g of 0.05 were detected. We report the balance of school- and pupil-level characteristics in Pupil and School Characteristics below.

Missing data analysis

The extent of missing data on the outcome and pre-treatment covariates was first explored descriptively using cross-tabulations, including counts and percentages in each category. We explored the extent of missingness and whether there is a pattern in missingness. A 'drop-out' model was estimated using a logistic regression to assess if there are patterns in missing data. The model included a binary outcome of whether primary outcome data or any primary analysis covariates are missing for each individual at follow-up, and all covariates outlined above in the imbalance at baseline analysis. If data is missing for these covariates this was coded as separate binary variables in the model. The model also included a random effect for schools. The 'drop-out' model was estimated using the `melogit` command in Stata 17. We followed 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 was employed. For more than 5% missing data overall, from baseline assessment to final analysis, our approach depended on the pattern of missingness. If the pattern of missingness was unrelated to the treatment effect (due to factors that affected testing but are not related to the programme, such as pupil absence due to illness) then missing data was assumed to be missing completely at random (MCAR) and we continued with a complete case analysis.

As data was observed to be missing in a way that is correlated with observable variables, the primary analysis was re-estimated through Multiple Imputation (MI) using Chained Equations (MICE). At the SAP stage, we had planned to use only significant variables to impute the primary outcome. However, we decided to include all variables used in the analytic model along with the auxiliary variables in the MI model to account for missing at random (MAR) and match the MI model to the analytic model, meeting the congeniality assumption (van Buuren and Groothuis-Oudshoorn, 2011; van Buuren and Oudshoorn, 2000; Woods et al., 2024). Considering the multilevel structure of the data, we additionally included school-level clusters as dummy indicators in the MI model, accounting for the association between the PTM outcome and the partially observed variables within clusters (Graham, 2012).[26] The imputed datasets were used to replicate the main analyses and compare the results with the complete data analysis.

Multiple imputation was conducted using the `mi` suite of commands in Stata 17. The process and results of imputation are reported in the Missing Data analysis below.

Compliance

The Complier Average Causal Effect (CACE)[27] was estimated to show the impact of Stop and Think on the primary outcome—maths attainment among KS2 FSM pupils—compared to individuals in the control group, taking into account the level of compliance with Stop and Think (see Analysis in the Presence of Non-Compliance).

Data for our compliance analysis was collected during the implementation period through the computer-assisted programme that delivers Stop and Think, which provided the number of sessions completed by each class (from 0 to 29)[28] used as a measure of if and how fully the intervention has been delivered to classes. Completed sessions are defined as sessions which were started and in which at least one question was completed. 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. Although the measures of compliance are at class level, the unit of analysis was pupils. Birkbeck shared the compliance data with BIT and NatCen.

[26] To maintain consistency, we carried out the multiple imputation in Stata. Although Stata does not support multilevel MI, it provides a few alternative approaches to enable us to carry out the MI model as compatibly as the analytic model. Two commonly used approaches are: a) including indicator variables for clusters in the MI model (the one we chose); and b) imputing data separately within each individual cluster (Graham, 2012). The second approach would not work for our data because one cluster had completely missing PTM outcomes due to attrition at the endline testing. We chose the first approach, especially given that we did not impute any level-two (i.e., school-level) variables and that the cluster sizes are not small, where a small cluster size might lead to biased estimates of ICCs, between-group coefficients, and standard errors (Luedtke et al., 2017; Woods et al., 2024).

[27] Corresponding to the average effect of the intervention for those pupils who have complied with the programme.

[28] There were 30 sessions to be completed by each class. However, the first session was an introductory video which did not involve any gameplay, so it was not tracked.

Further, considering that school-level unobserved characteristics might influence both compliance with the intervention and the primary outcome, we estimated the CACE using a two-stage least square (2SLS) model (Angrist and Imbens, 1995) with treatment allocation as the instrumental variable (IV) for the compliance measure.

As the first stage of the model, compliance was regressed on all covariates that are used in the main primary outcome model, including additionally a binary variable that indicates a pupil's pre-intervention treatment allocation as an IV. 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 regressed the primary outcome on the covariates used in the main models and also included 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:

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

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 is equal to the intention-to-treat estimate.

As use of the Stop and Think software was not restricted after the end of the intervention period, we conducted two sets of compliance analysis. We first explored the number of sessions completed before the intervention end date. We then also provide an additional robustness check, carrying out the compliance analysis using the total number of sessions completed before endline testing. We hence had two compliance measures: (a) compliance truncated to the intervention end date and (b) compliance untruncated until the date of endline testing.

IV regression was conducted in Stata 17 using the `ivregress` command and the `cluster` option to control for clustering on schools.

Estimation of intra-cluster correlations (ICCs)

The intra-cluster correlations (ICCs) were estimated directly from the primary analysis model, using the variance estimates for each level of clustering. The ICC for schools $\boldsymbol{\rho S}$ was estimated with the post-estimation command `estat icc` in Stata 17, using the following formula based on Hedges (2011):

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

where σ^2_{BS} is the between-school variance, σ^2_{WS} is the within-school variance and σ^2_{WT} is the total variance.

Estimation of effect sizes

Effects size calculation for primary and secondary outcome analyses

Effect sizes (ES) for cluster-randomised trials were used, as adapted from Hedges (2007):

$$ES = \frac{(\underline{Y_T} \; - \underline{Y_C}\;)_{adjusted}}{\sqrt{\sigma^2_u + \sigma^2_e}}$$

where

$(\underline{Y_T} \; - \underline{Y_C}\;)_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics;

$\sqrt{\sigma^2_u + \sigma^2_e}$ 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 took each group's adjusted mean and variance to calculate the effect size. The variance was 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 accounts the clustering, is also reported.

Effects size calculation for the sensitivity analysis for primary outcome

As mentioned, we 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.[29] We will use the ESs for cluster-randomised trials, as adapted from Hedges (2007):

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

where

$(\underline{Y_T} - \underline{Y_C})_{adjusted}$ is the mean difference between the intervention and control group adjusted for baseline characteristics;

$\sqrt{\sigma_u^2 + \sigma_e^2}$ 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 took each group's adjusted mean and variance to calculate the effect size. The variance was 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, is also 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.

Similarly to Roy et al. (2019), we followed the method described by Borenstein et al. (2009, p.218) to obtain the combined effect size. To obtain this, we first calculated the weights assigned in each model:

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

where, V_3 and V_5 are variances for the Year 3 and Year 5 models, respectively. The combined effect size was then calculated as:

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

Lastly, the combined variance was calculated as:

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

Effect size calculation for further sensitivity analysis

As above, we used the ESs for cluster-randomised trials, as adapted from Hedges (2007):

[29] 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

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

where

$(AttainmentGap)_{adjusted}$ (i.e., β_4 as per in the model above) is the difference in average effect of the intervention between FSM pupils and their peers adjusted for baseline characteristics;

$\sqrt{\sigma_u^2 + \sigma_e^2}$ 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, is also reported.

As a sensitivity check and for comparability of ES between the interaction models and that derived from the main analyses of FSM-eligible subsample (RQ1 and RQ4), we further followed the EEF statistical analysis guidance (EEF, 2022) to calculate the ES using the following equation:

$$ES_{FSM\ subgroup} = \frac{\beta_2 Intervention_j + \beta_4 Intervention_j FSM_{ij}}{sd_{FSM}}$$

where β_2 and β_4 correspond to the specifications outlined in the Further Sensitivity Analysis section and sd_{FSM} is unconditional standard deviation of the FSM-eligible subsample. Ideally, since the main analyses used a restricted sample of pupils eligible for FSM, the ES extracted from the interaction models should be analogous to that from the main models (RQ1 and RQ4), despite using different calculation approaches.

Implementation and process evaluation

Research methods

IPE research activities were conducted in three phases, as illustrated in Figure 4. This approach ensured the timing of activities built on one another and minimised the burden on schools.

Figure 4: Phases of IPE research activities

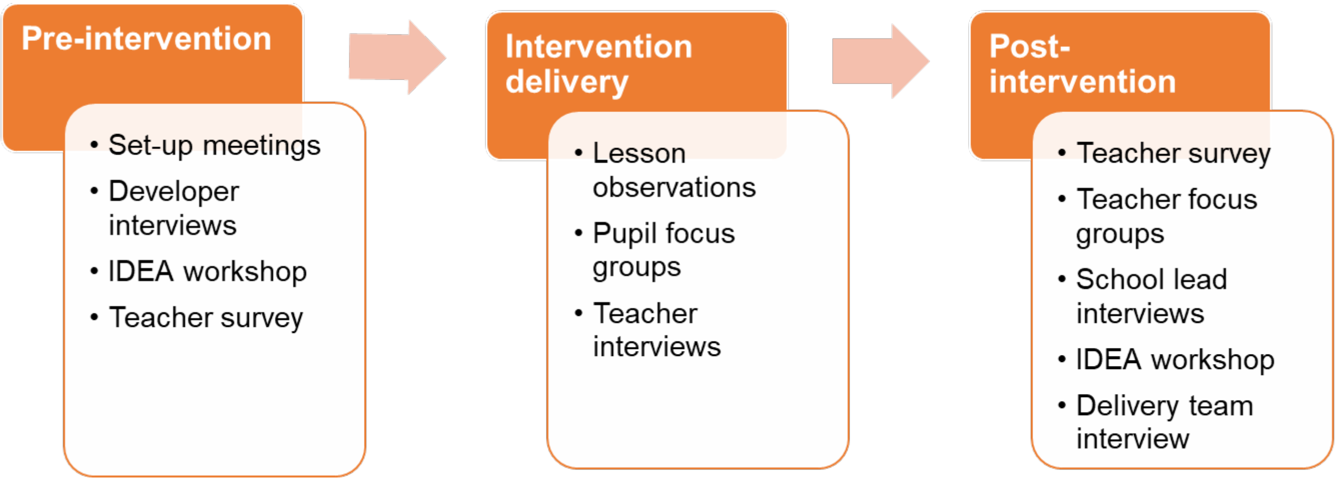


Table 8 gives an overview of IPE research methods and how these related to the research questions.

Table 8: IPE methods overview

Research methods	Data collection methods	Data analysis methods	Research questions addressed	Implementation/logic model relevance
Developer interviews	Telephone/online	Qualitative	1, 3	Fidelity, adaptation
Delivery team interview	Online	Qualitative	1, 2	Fidelity, quality
Lesson observations	Face to face	Qualitative	1, 2, 3, 6	Fidelity, quality, adaptation, responsiveness
Pupil focus groups	Face to face	Qualitative	6, 7, 9	Responsiveness, differentiation, perceived impacts
Teacher interviews	Face to face, telephone, and online	Qualitative	1, 2, 5, 6, 7, 9	Responsiveness, differentiation, perceived impacts
Teacher surveys (pre and post intervention)	Online	Quantitative	1, 3, 4, 5, 6, 7, 9	Fidelity, adaptation, dosage, reach, responsiveness, differentiation, perceived impacts
Teacher focus groups	Online	Qualitative	1, 3, 5, 6, 8	Fidelity, adaptation, responsiveness, differentiation, impacts
School lead interviews	Online	Qualitative	3, 6, 7, 8	Adaptation, responsiveness, differentiation, monitoring the control

Pre-intervention

Pre-intervention data collection identified the nature and reasons for changes to the intervention from the efficacy trial and teacher attitudes and usual practice teaching methods before Stop and Think.

In set-up meetings, and an Intervention Development and Evaluation Analysis (IDEA) workshop with Birkbeck, BIT, and the EEF in autumn term 2020, we reviewed the logic model used in the efficacy trial and identified anticipated changes to the software or delivery model. This included exploring the implications of a new delivery partner (BIT, rather than Birkbeck as in the efficacy trial). The logic model was used to finalise primary and secondary pupil-level outcomes and informed our examination of the IPE dimensions.

In October 2020, we interviewed the lead developer at Birkbeck to map out plans, explore key learning from the efficacy trial, and identify any anticipated delivery challenges. In November 2022, we conducted a second interview with the lead developer. This took place following the pause to delivery and evaluation activities due to the COVID-19 pandemic. The second interview explored the plans and expectations from intervention delivery, any changes made to the programme, and any anticipated delivery challenges.

In the autumn term 2022, classes were informed of their randomisation allocation. This was followed by a survey of all teachers allocated to treatment and control in the autumn term 2022, immediately prior to the intervention period, to gather quantitative data on teaching as usual in maths and science and attitudes about, and experiences of, using technology in education. The survey also explored understanding and expectations of the intervention.

Headteachers, assistant headteachers, and science leads acted as the school leads and point of contact for the schools for the Stop and Think evaluation. We estimate that 593 teachers participated in the overall trial based on the number of classes, as shown in Table 7. It is important to note that some schools had more than one Year 3 and Year 5 class and it is also likely that some classes had additional TAs involved in the delivery of Stop and Think. Additionally, of the estimated 593 classes, 294 were in the control group so we cannot be sure how engaged they remained with the trial after their randomisation allocation was shared with them.

The survey link was sent to 172 school leads and they were asked to share the link with Year 3 and Year 5 teachers. Because school leads forwarded the survey link on our behalf, we cannot ascertain the total number of teachers that received this link from their school leads. This meant that we are unable to provide percentage response rates for the survey since we do not have a fixed total sample.

We received 157 responses to the survey: 107 from treatment teachers and 41 from control teachers. Nine survey respondents were unsure about whether they were in the treatment group or the control group.

During intervention delivery

IPE activities during the delivery period comprised school visits to gather 'in-the-moment' reflections on the acceptability of Stop and Think, fidelity of implementation, reasons for any adaptation, and early reflections of perceived impacts.

We conducted four school visits—one each for a Year 3 maths lesson, Year 3 science lesson, Year 5 maths lesson, and Year 5 science lesson. Each visit began with a lesson observation followed by a pupil focus group (with pupils who had been in the observed lesson) and finally an interview with the teacher who had delivered the lesson. These visits took place in the middle of the ten-week delivery period, with each school visit taking place during a different delivery week: weeks three, four, five, or six. We strived to ensure this variation during the school visits and worked with schools to schedule visits to suit their availability.

The lesson observations assessed fidelity of implementation and teachers' and pupils' responsiveness. The data gathered was used to tailor the prompts and probes used to inform the subsequent pupil focus group and teacher interviews.

We conducted one pupil focus group in each of the four schools to explore responsiveness and understand whether, and how, the programme was encouraging pupils to give slower and more reflective answers. We explored occasions when pupils felt they had 'stopped and thought' where previously they would not have. During the intervention, pupils were explicitly told that the programme aims to help them stop and think. Pupil focus groups therefore looked for evidence of explicit understanding as an intended mechanism through which change occurs.

We interviewed one teacher in each school who was delivering the programme to explore responsiveness, the extent to which the intervention was being delivered as intended (including any adaptation or tailoring and whether this was useful), and how different the programme is to their usual teaching. We investigated how Stop and Think is incorporated into planning, any support (including technical assistance) required or received, and the practicalities of delivering the intervention in different school or class contexts. We also gathered teachers' reflections on early perceived impacts.

Post-intervention

Post-intervention IPE research activities took place late in summer term 2023 and comprised a teacher survey, focus groups with teachers, and interviews with school leads. These activities focused on understanding adaptation and variability at scale as well as exploring perceived impacts.

We conducted a post-intervention survey of all teachers allocated to treatment and control to gather quantitative data on fidelity, tailoring, and adaptation at scale. The survey explored barriers and facilitators to successful delivery and any perceived impacts on intermediate outcomes. It also included usual practice and differentiation questions for control group teachers. We received 118 responses: 69 from treatment teachers, and 49 from control teachers.[30] As with the pre-intervention survey, we are unable to provide response rates for the survey as we do not know the total number of teachers that received the survey link. The link was shared with 176 school leads (some schools appointed more than one school lead) who were asked to share it with Year 3 and Year 5 teachers.

We conducted four focus groups with teachers who had delivered the Stop and Think programme to explore variation in delivery and identify any reasons for adaptation. In total, the focus groups covered 13 teachers—seven from Year 3 and six from Year 5. The focus group methodology facilitated teachers reflecting on their practice more critically through hearing from others at different schools and comparing their approach with them.

We interviewed seven Stop and Think school leads who had responsibility for liaising between the delivery team and class teachers and overseeing the implementation of the programme in their school. The sample consisted of headteachers, maths or science subject leads, and teachers who were delivering the programme. The interviews gathered data on any adaptations made to delivery, responsiveness, and teaching as usual in the year group allocated to treatment, and monitoring of the control. We also explored barriers and facilitators to delivery, how the intervention was incorporated into lesson or curriculum planning, and the school's motivations for taking part in the evaluation.

Post intervention, we had a final interview with the delivery lead at BIT to explore any adaptations made to delivery guidance, barriers and facilitators to school engagement, perceived outcomes for teachers, and explore recommendations for improvements to the intervention.

Finally, we analysed software data on the number and spacing of sessions to assess fidelity (whether schools delivered Stop and Think sessions according to the intended schedule—three times per week over a ten-week period) and dosage (the number of Stop and Think sessions delivered). Compliance analysis was conducted as part of the impact evaluation.

IPE data collection

We developed all research tools with reference to the finalised logic model and delivery plans, findings from the efficacy trial, and earlier research activities. This enabled us to triangulate perspectives and verify emerging findings.

During recruitment, and before all IPE data collection activities, the research team explained to participants that we are independent evaluators (operating separately from the teams at Birkbeck and BIT) and gathered informed consent.

We conducted interviews with education professionals online or by telephone according to participant preference. This approach allowed us to be responsive and accommodating to teachers' timetables and the nature of the school day. The teacher focus groups were also conducted online, allowing participants from across the country to meet together.

The lesson observations and pupil focus groups were conducted by researchers experienced in conducting research with children. Interviews and teacher focus groups lasted up to an hour and the pupil focus groups no more than 50 minutes—the shorter time for these being more suited to attention span of primary school children. All data collection activities were audio recorded with participant permission.

Both the pre- and post-intervention teacher surveys were administered online and took up to 15 minutes to complete.

IPE recruitment and sampling

Qualitative methods

At each phase of the qualitative IPE research activity, schools were sampled to achieve diversity in characteristics expected to affect teacher and pupil experiences of Stop and Think, and the way it is incorporated into planning and teaching, for example, FSM eligibility, type of school (maintained or academy), school size, and location of the school. Table 9 shows the achieved samples against the sample aim for each phase of the IPE qualitative activity.

Table 9: IPE sample targets and achieved samples

Data collection method	Sample target	Sample achieved
Lesson observations	4 lesson observations	4 lesson observations
Pupil focus groups	4 groups c. 6 pupils per group	4 groups 23 pupils across 4 groups
Teacher interviews	8 interviews	4 interviews
Teacher focus groups	4 groups c. 5 teachers per group	4 groups 13 teachers across 4 groups[31]
School lead interviews	8 interviews	7 interviews

School visits

A purposive sampling strategy was used to select four schools. First, we sampled for two Year 3 treatment group schools and two Year 5 treatment group schools. Second, we sampled for setting type (academies and maintained schools), FSM rate (reported by schools during trial recruitment), and school size.

For each of the schools visited, the intervention teacher from the lesson observation took part in the teacher interview and also selected the pupils to take part in the focus group. Teachers were asked to select a mix of boys and girls and pupils with a mixture of abilities. These school visits took place between February 2023 and May 2023.

School lead interviews

The post-intervention sampling strategy was staged. The first stage involved contacting a sample of 60 schools based on the primary criterion of pupils eligible for FSM: 20 schools with a low proportion of FSM pupils (less than 20%), 20 with a medium proportion (20% to 40%), and 20 with a high proportion (more than 40%). The secondary sampling criteria used were:

- school status—both academy and maintained schools;
- school size—both smaller and larger schools;
- treatment year group—Year 3 and Year 5 teachers; and
- geographical spread—regions from across England.

[31] The focus group recruitment aimed at recruiting five teachers per group, however, some scheduled groups experienced last-minute withdrawals or no-shows.

We liaised with school leads to find convenient times for online interviews. School lead interviews were conducted between May and July 2023.

Teacher focus groups

Recruitment for the four focus groups proved challenging and required widening the net from a purposive sample to include all schools. School leads were asked to share the focus group invitation with the relevant teachers at their school. The teacher focus groups were carried out between May and July 2023. The participants came from schools across the sample including from:

- schools with low, medium, and high FSM eligibility, ranging from 3.5% to 49.6%;
- schools in different regions, from North East to South West of England;
- different sized schools, from a very small school with 13 pupils in the year group to larger schools with four classes per year group; and
- a mixture of both maintained and academy schools.

Quantitative methods

For the pre- and post-intervention teacher surveys, all teachers taking part in the intervention were invited to complete them. We asked the Stop and Think school leads to circulate the survey to their teachers. The pre-intervention survey was conducted between November and December 2022 and the post-intervention survey between May and July 2023.

One hundred and fifty-seven teachers completed the pre-intervention survey—107 from the treatment group and 41 from control. The post-intervention survey was completed by 118 respondents—69 from the treatment group and 49 from the control group.

Analysis

We managed and analysed qualitative data using the Framework approach developed by NatCen.[32] Framework is a systematic matrix approach that allows analysis within and across cases and themes. Using themes covered in topic guides and any other themes which emerged from the data, we assembled a matrix in which each row represents an individual interview or focus group discussion and each column a theme and any related sub-themes. We then summarised the qualitative data in the matrix, including illustrative verbatim quotes where appropriate. Once all data was managed in this matrix we moved to formal analysis. This involved a first phase of familiarisation with the dataset moving to a phase of 'detection', including studying the elements participants say about a particular phenomenon, listing these and sorting them thematically in relation to the research questions. Once we identified different themes in the data we created higher-level categories that work as meaningful conceptual groupings for participants' views and experiences.

We managed and analysed the survey and software data using Excel. Our analysis explored differences in IPE domains between Year 3 and Year 5, where sample sizes allowed.

We triangulated and synthesised IPE data according to the research questions and implementation domains. This enabled us to provide a comprehensive assessment of implementation, report against the finalised logic model, and explain the impact evaluation findings. Our analysis also draws out key learning for future delivery, including any potential changes required for future scale-up of the intervention.

[32] Ritche et al. (2013) *Qualitative Research Practice*, London: Sage.

Costs

We collected and analysed cost data from both schools and the delivery partner for the intervention (BIT) in line with EEF guidelines (EEF, 2023).

To obtain information on the costs of taking part in Stop and Think for schools, cost-related questions were included in the teacher survey part of the IPE and distributed at endline. The questions explored the categories of personnel costs for training and implementation of the programme and included teacher time for the following intervention-related activities:

- teacher cover required while in training;
- time spent preparing for the first session;
- time spent preparing for subsequent sessions; and
- extra time spent delivering lessons when using Stop and Think.

The survey was completed by 118 teachers across conditions. In line with EEF guidance, our analysis focuses on the *additional* cost incurred by exploring responses of the teachers who delivered Stop and Think on the questions above. The results we report (for example, mean time spent) are based on non-missing responses for each item unless otherwise detailed.

In addition to the costs collected from schools, a form was sent to BIT to collect information on delivery partner costs. The form covered the following categories:

- personnel costs for preparing programme delivery;
- personnel costs during training for implementation of the programme;
- personnel costs for the implementation of the programme; and
- facilities, equipment, and materials for implementation.

Data from schools and BIT was used to calculate the per-pupil cost of implementation over three years (EEF 2023), categorising costs into pre-requisites, start-up, and recurring costs. Per-pupil cost estimates were based on the number of pupils enrolled in the trial across 173 schools (n = 14,718).

Timeline

Table 10: Stop and Think evaluation timeline

Dates	Activity	Staff responsible or leading
July–September 2020	Initial set up meetings	EEF, Birkbeck, BIT, NatCen
October–December 2020	IDEA workshop; ethical approval	NatCen
December 2020	Developer interview	NatCen
December 2020	Recruitment materials developed	BIT, NatCen
May 2021	Trial registered on ISRCTN	NatCen
September–October 2021	Set-up meeting with the EEF	EEF, Birkbeck, BIT, NatCen
September–October 2021	Recruitment materials updated	BIT, NatCen
October–July 2022	School recruitment	BIT
February 2022	Protocol published	NatCen

July 2022	School sample shared with NatCen	BIT
September–October 2022	Pupil enumeration; NPD application; developer interview	NatCen
October 2022	Randomisation	NatCen
November–December 2022	Pre-intervention survey	NatCen
October 2022–January 2023	Training sessions for teachers	BIT
January 2023	Statistical Analysis Plan published	NatCen
February–May 2023	Intervention running in schools	BIT
February–May 2023	IPE school visits	NatCen
May–July 2023	Post-intervention IPE data collection Endline testing	NatCen
September 2023–January 2025	Analysis and drafting	NatCen
March 2025	Publication of report final	NatCen

Impact evaluation results

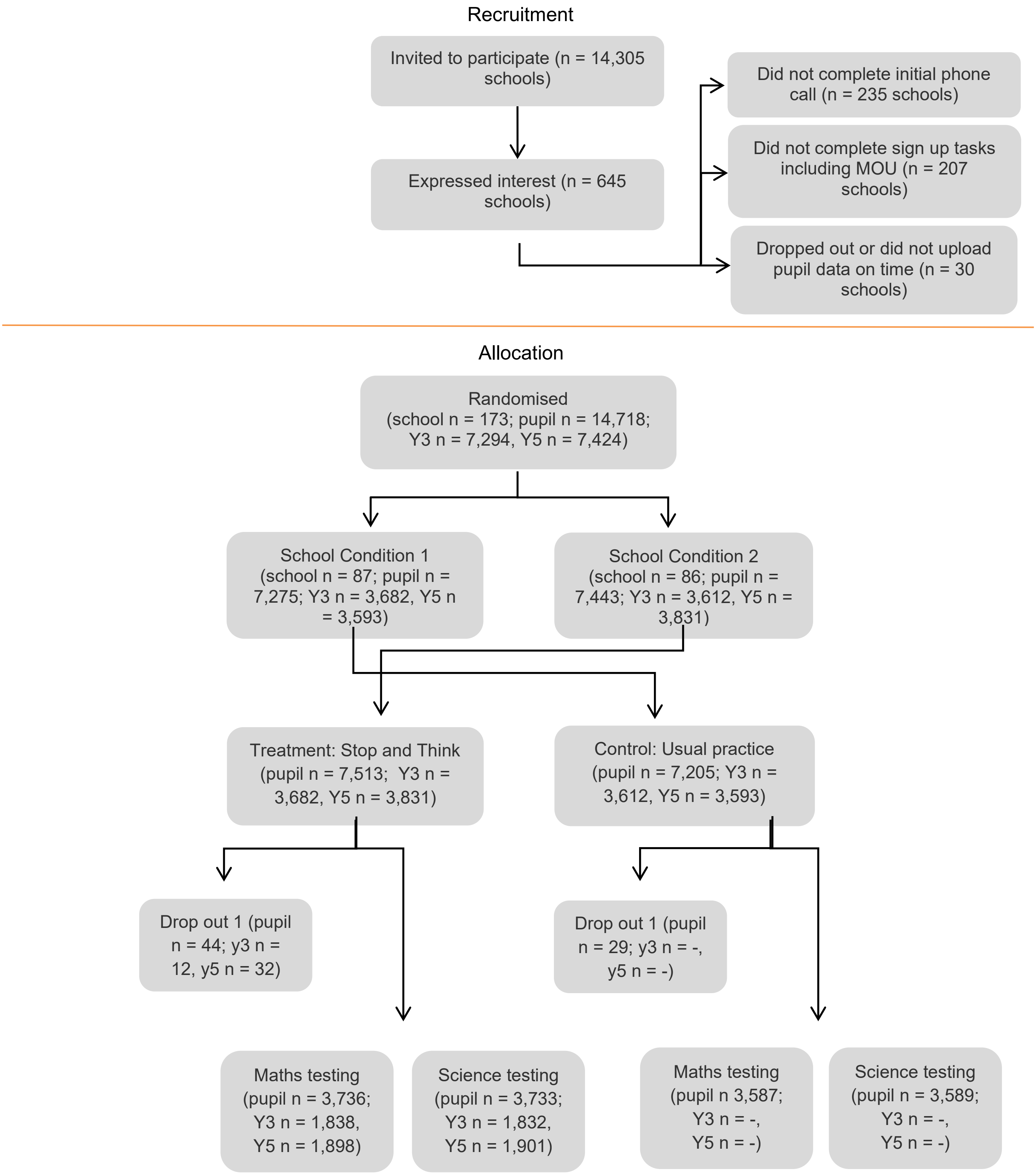
Participant flow including losses and exclusions

Figure 5, Figure 6, and Figure 7 present the participant flow diagram from recruitment to analysis by subject. A total of 173 schools were recruited and randomised into the two school conditions (Condition 1: Year 3 receives the intervention and Year 5 continues teaching as usual; Condition 2: Year 5 receives the intervention and Year 3 continues teaching as usual). This meant that a total of 7,513 pupils were randomised to Stop and Think and 7,205 pupils were randomised to teaching as usual. As this analysis is subject to the SRS statistical disclosure guidance, we are not displaying some counts to avoid statistical disclosure.

Three schools (73 pupils) withdrew from the trial between treatment allocation and the randomisation of pupils to sit either maths or science tests. This meant that 14,645 pupils (7,469 treatment and 7,176 control) in 170 schools were randomly allocated to maths or science testing. We obtained consent to process NPD data from 14,587 pupils (7,439 treatment and 7,148 control) across 169 schools.

A total of 594 pupils were excluded from the sample for analysis due to missing baseline NPD data. We excluded a further 1,916 pupils from the primary and secondary analyses on the basis that baseline data was available on the NPD but that they did not complete endline testing. The total sample available for analysis therefore comprised 12,077 pupils across 167 schools. More details can be found in the Attrition section below.

Figure 5. Participant flow diagram—recruitment to random testing allocation


Recruitment
Invited to participate (n = 14,305 schools)
Expressed interest (n = 645 schools)
Did not complete initial phone call (n = 235 schools)
Did not complete sign up tasks including MOU (n = 207 schools)
Dropped out or did not upload pupil data on time (n = 30 schools)
Allocation
Randomised (school n = 173; pupil n = 14,718; Y3 n = 7,294, Y5 n = 7,424)
School Condition 1 (school n = 87; pupil n = 7,275; Y3 n = 3,682, Y5 n = 3,593)
School Condition 2 (school n = 86; pupil n = 7,443; Y3 n = 3,612, Y5 n = 3,831)
Treatment: Stop and Think (pupil n = 7,513; Y3 n = 3,682, Y5 n = 3,831)
Control: Usual practice (pupil n = 7,205; Y3 n = 3,612, Y5 n = 3,593)
Drop out 1 (pupil n = 44; y3 n = 12, y5 n = 32)
Drop out 1 (pupil n = 29; y3 n = -, y5 n = -)
Maths testing (pupil n = 3,736; Y3 n = 1,838, Y5 n = 1,898)
Science testing (pupil n = 3,733; Y3 n = 1,832, Y5 n = 1,901)
Maths testing (pupil n 3,587; Y3 n = -, Y5 n = -)
Science testing (pupil n = 3,589; Y3 n = -, Y5 n = -)

Note: '-' in this figure denotes suppression to avoid statistical disclosure, in accordance with DfE SRS statistical disclosure guidance.

Figure 6. Participant flow diagram—maths testing allocation to analysis

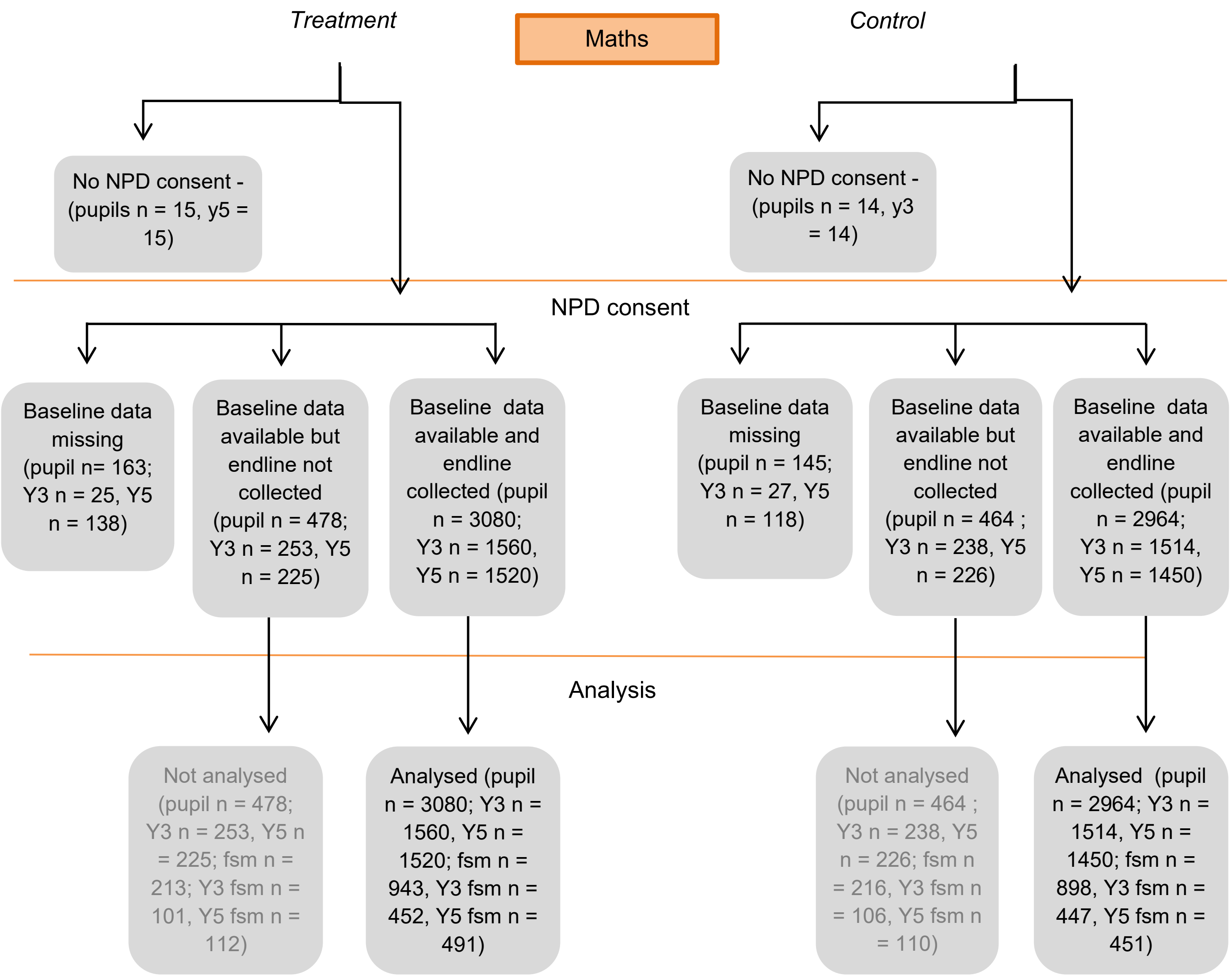
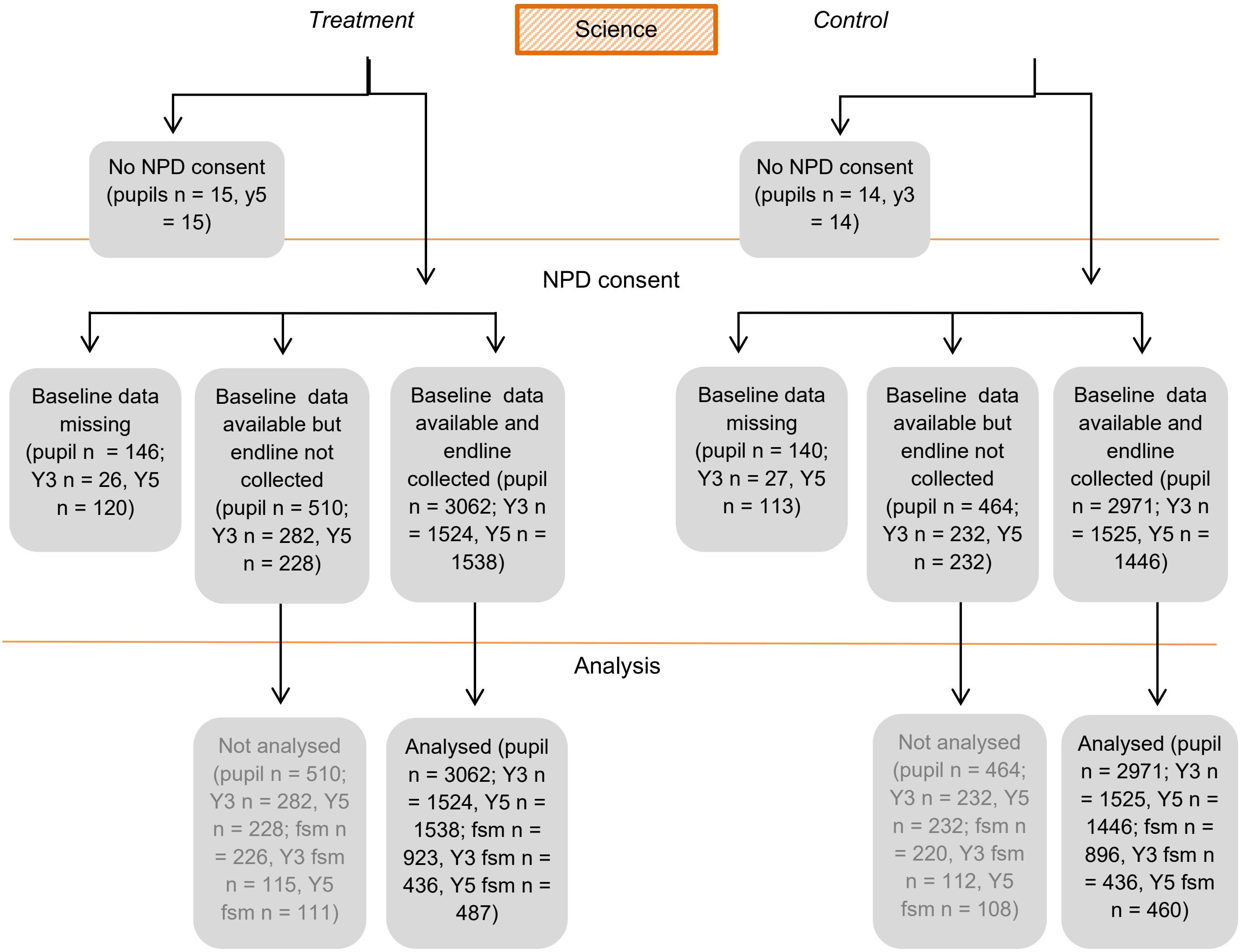

Treatment
Maths
Control
No NPD consent - (pupils n = 15, y5 = 15)
No NPD consent - (pupils n = 14, y3 = 14)
NPD consent
Baseline data missing (pupil n= 163; Y3 n = 25, Y5 n = 138)
Baseline data available but endline not collected (pupil n = 478; Y3 n = 253, Y5 n = 225)
Baseline data available and endline collected (pupil n = 3080; Y3 n = 1560, Y5 n = 1520)
Baseline data missing (pupil n = 145; Y3 n = 27, Y5 n = 118)
Baseline data available but endline not collected (pupil n = 464 ; Y3 n = 238, Y5 n = 226)
Baseline data available and endline collected (pupil n = 2964; Y3 n = 1514, Y5 n = 1450)
Analysis
Not analysed (pupil n = 478; Y3 n = 253, Y5 n = 225; fsm n = 213; Y3 fsm n = 101, Y5 fsm n = 112)
Analysed (pupil n = 3080; Y3 n = 1560, Y5 n = 1520; fsm n = 943, Y3 fsm n = 452, Y5 fsm n = 491)
Not analysed (pupil n = 464 ; Y3 n = 238, Y5 n = 226; fsm n = 216, Y3 fsm n = 106, Y5 fsm n = 110)
Analysed (pupil n = 2964; Y3 n = 1514, Y5 n = 1450; fsm n = 898, Y3 fsm n = 447, Y5 fsm n = 451)

Figure 7. Participant flow diagram—science testing allocation to analysis


Treatment
Science
Control
No NPD consent (pupils n = 15, y5 = 15)
No NPD consent (pupils n = 14, y3 = 14)
NPD consent
Baseline data missing (pupil n = 146; Y3 n = 26, Y5 n = 120)
Baseline data available but endline not collected (pupil n = 510; Y3 n = 282, Y5 n = 228)
Baseline data available and endline collected (pupil n = 3062; Y3 n = 1524, Y5 n = 1538)
Baseline data missing (pupil n = 140; Y3 n = 27, Y5 n = 113)
Baseline data available but endline not collected (pupil n = 464; Y3 n = 232, Y5 n = 232)
Baseline data available and endline collected (pupil n = 2971; Y3 n = 1525, Y5 n = 1446)
Analysis
Not analysed (pupil n = 510; Y3 n = 282, Y5 n = 228; fsm n = 226, Y3 fsm n = 115, Y5 fsm n = 111)
Analysed (pupil n = 3062; Y3 n = 1524, Y5 n = 1538; fsm n = 923, Y3 fsm n = 436, Y5 fsm n = 487)
Not analysed (pupil n = 464; Y3 n = 232, Y5 n = 232; fsm n = 220, Y3 fsm n = 112, Y5 fsm n = 108)
Analysed (pupil n = 2971; Y3 n = 1525, Y5 n = 1446; fsm n = 896, Y3 fsm n = 436, Y5 fsm n = 460)

Attrition

Across the 169 schools for which NPD consent was obtained, data was processed for 7,439 treatment pupils and for 7,148 control pupils. In NPD records obtained, baseline data was not available for 309 treatment and 285 control pupils.[33] Among these were (a) 77 treatment and 75 control pupils for whom baseline data was not available and no endline data was collected and (b) 232 treatment and 210 control pupils for whom baseline data was not available but endline data was collected.

Baseline data was available for 7,130 treatment and 6,863 control pupils. Of these, for 988 treatment and 928 control pupils endline data was not collected. This is due to the one school not proceeding to endline testing but consenting for NPD data to be processed, as well as a number of pupils across schools not completing endline testing due to, for example, being absent on the day of testing. Baseline and endline data were hence available and analysed for 6,142 treatment and 5,935 control pupils.

In total, 1,371 treatment and 1,270 control pupils were lost from treatment allocation to analysis, representing a 18.25% attrition rate in the treatment group and a 17.63% attrition rate in the control group. **Error! Reference source not found.**1 presents pupil-level attrition from the trial.

At the school level, six schools were lost from condition allocation to analysis representing a 3.47% school attrition rate, as presented in **Error! Reference source not found.**.

[33] While there is no difference, on average, in the number pupils whose baseline information is missing between treatment and control groups across all school, there are variations between schools.

Table 11: Pupil-level attrition from the trial—overall and primary outcome

		Overall			**Maths**			**Science**			**Maths + FSM (primary outcome)**		
		Total	**Treatment**	**Control**	**Total**	**Treatment**	**Control**	**Total**	**Treatment**	**Control**	**Total**	**Treatment**	**Control**
Number of pupils	Treatment allocation	14,718	7,513	7,205									
	Test subject allocation	14,645	7,469	7,176	7,323	3,736	3,587	7,322	3,733	3,589			
	NPD consent	14,587	7,439	7,148	7,294	3,721	3,573	7,293	3,718	3,575	2,375	1,215	1,160
	Analysed	12,077	6,142	5,935	6,044	3,080	2,964	6,033	3,062	2,971	1,841	943	898
Pupil attrition (treatment allocation to NPD consent)	Number	131	74	57									
	Percentage	0.89%	0.98%	0.79%									
Pupil attrition (NPD consent to analysed)	Number	2,510	1,297	1,213	1,250	641	609	1,260	656	604	534	272	262
	Percentage	17.21%	17.44%	16.97%	17.14%	17.23%	17.04%	17.28%	17.64%	16.90%	22.48%	22.39%	22.59%
Pupil attrition (treatment allocation to analysed)	Number	2,641	1,371	1,270									
	Percentage	17.94%	18.25%	17.63%									

Table 12. School-level attrition from the trial—overall and primary outcome

		Overall			Maths + FSM (primary outcome)		
		Total	Condition 1	Condition 2	Total	Condition 1	Condition 2
Number of schools	Condition allocation	173	87	86			
	Test subject allocation	170	86	84			
	Retained NPD data	169	86	83	165	83	82
	Analysed	167	85	82	164	83	81
School attrition (treatment allocation to NPD consent)	Number	4	1	3			
	Percentage	2.31%	1.15%	3.49%			
School attrition (treatment allocation to analysed)	Number	6	2	4			
	Percentage	3.47%	2.30%	4.65%			

Pupil and school characteristics

To assess the balance of pupil and school characteristics at baseline, we present descriptive analysis at pupil and school levels using all data available for schools that gave NPD consent. Data on key school and pupil characteristics (pupil baseline scores, FSM eligibility in the previous six years) was not available at randomisation, hence schools that withdrew between randomisation and NPD consent are included in this analysis as missing data on those characteristics.

Table 13 displays pupil characteristics across the Stop and Think and control groups and school characteristics across the two school conditions (condition 1: Year 3 receives the intervention and Year 5 continues teaching as usual; condition 2: Year 5 receives the intervention and Year 3 continues teaching as usual). At the school level, the proportion of pupils eligible for FSM at any point in the previous six years was similar between school conditions, with schools in Condition 2 having a near identical proportion on average. Across both conditions, the highest proportion of schools had a class form entry size of one, followed by a class form entry size of two. Schools in Condition 1 were slightly more likely to be of class form entry size one, compared to schools in Condition 2. At the pupil level, FSM eligibility in the previous six years was equal between the Stop and Think and control groups at a third of pupils. Average standardised baseline scores were slightly higher in the control group compared to the Stop and Think group with the number of missing baseline scores slightly higher in the latter.[34]

As specified in the Methods section above, an effect size of greater than 0.05 was considered an indication of possible imbalance. Given that the effect size was not greater than 0.05 for either variable, we did not conduct a sensitivity analysis including unbalanced variables in the models.

Table 13. Baseline characteristics of groups at randomisation

School-level (categorical)	**Condition 1**		**Condition 2**		
	n/N (missing)	**Count (in %)**	**n/N (missing)**	**Count (in %)**	
Class form entry size: 1	38/87 (1)	44%	36/86 (3)	42%	
Class form entry size: 2	32/87 (1)	37%	32/86 (3)	37%	
Class form entry size: 3	16/87 (1)	18%	15/86 (3)	17%	
School-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size (Hedge's g)**
School-level FSM	86/87 (1)	32.07 (17.32)	83/83 (3)	32.99 (19.89)	-0.05
Pupil-level (categorical)	**Stop and Think**		**Teaching as usual (control)**		
	n/N (missing)	**Count (in %)**	**n/N (missing)**	**Count (in %)**	
FSM eligibility in the past six years: Yes	2,408/7,513 (107)	32.5%	2,319/7,205 (83)	32.5%	
Pupil-level (continuous)	**n/N (missing)**	**Mean (SD)**	**n/N (missing)**	**Mean (SD)**	**Effect size (Hedge's g)**
Baseline score (standardized)	7,130/7,513 (383)	-0.02 (1.03)	6,863/7,205 (342)	0.02 (0.97)	0.05

[34] As mentioned in the Method section, we used KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils due to COVID-19 pandemic interruption. Given the different scale of baseline measures for Year 3 and Year 5 pupils, we standardised the baseline measure to have a mean of zero and standard deviation of one by each year group. We then combined them to form a single baseline measure.

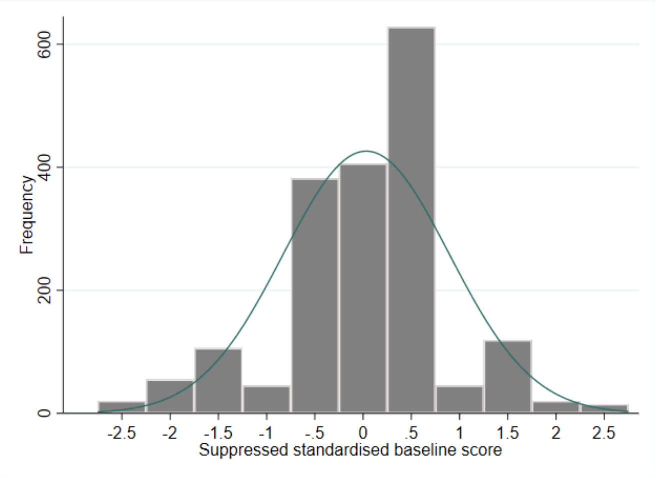
Outcomes and analysis

Primary analysis

The primary analysis explored the impact of Stop and Think on maths attainment of Year 3 and Year 5 pupils from disadvantaged backgrounds, as measured by FSM status (RQ1). The primary analysis used KS1 maths scores for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils as the baseline measure.

To mitigate the risk of multicollinearity, where the baseline measure is correlated with year group, we standardised the baseline measure to have a mean of zero and standard deviation of one by each year group. The FSM pupils in the final primary analysis had a mean of -0.17 for the standardised baseline measure (SD = 92; Skewness = -0.60; Kurtosis = 4.49).[35] Following the SRS statistical disclosure guidance about presenting graphic representations of data, we grouped the standardised baseline measure into 11 categories, with -2.5 representing a score below or equal to -2.5, -2 for those greater than -2.5 but below or equal to -2, and so on. This is to ensure that the underlying counts meet or exceed the threshold of ten. Figure 8 presents the grouped standardised baseline score, with 34% (n = 628) of pupils obtaining a standardised baseline score greater than zero but below or equal to 0.5. Following the disclosure guidance, we also do not report minimum or maximum values of all measures at baseline and endline. We instead grouped the measures and present graphic representations of data to ensure the minimum or maximum values are shared by at least ten individuals.

Figure 8: Distribution of grouped standardised baseline score for primary analysis



The baseline score refers to z-score standardised baseline measure (KS1 maths scores for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils).

The primary analysis used the age-standardised scores from GL Assessment's Progress Tests in Maths (GL PTM) as the primary outcome for FSM-eligible pupils. The FSM pupils in the final analysis (1,841 pupils with observable data on all variables) had an overall mean of 89.48 and standard deviation of 13.87 for the age-standardised GL PTM at endline testing (Skewness = 0.47; Kurtosis = 2.76).[36] The intervention group had a mean of 89.64 and standard deviation of 13.91 (Skewness = 0.51; Kurtosis = 2.80) for the age-standardised GL PTM at endline testing while the control group had a mean of 89.31 and standard deviation of 13.83 (Skewness = 0.43; Kurtosis = 2.70).

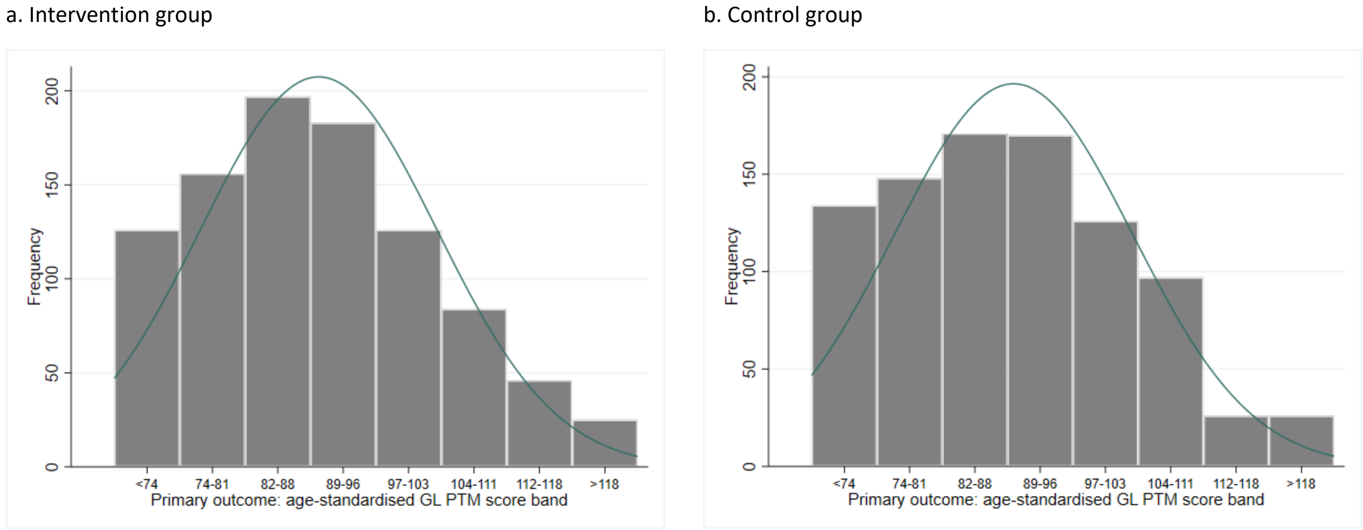
The primary outcome measure was moderately correlated with the standardised baseline measure with r = 0.51 (p < 0.001) among FSM-eligible pupils. Here, again, to adhere to SRS statistical disclosure guidance, we grouped the age-standardised GL PTM score into nine categories for presenting data, which is in line with the GL Assessment

[35] The skewness is a measure of the asymmetry of a distribution where a value of zero indicates that there is no skewness in the distribution (i.e., symmetrical). Kurtosis is a measure of whether or not a distribution is heavy-tailed or light-tailed relative to a normal distribution, where a kurtosis less than three suggests that the data has fewer extreme outliers than a normal distribution.

[36] The statistics suggest that the distribution of age-standardised GL PTM at endline among FSM pupils was symmetrical, though light-tailed.

standard age score bands.[37] Figure 9 illustrates the distribution of the primary outcome—age-standardised GL PTM score bands at endline by group (the overall distribution is presented in Figure 21 in Appendix D). The figure shows a peak at the band of 82–88, with around 20% (intervention n = 197; control n = 171) of pupils obtaining this score band. Due to the non-normal distribution of the dependent variable, residuals were plotted against fitted values to assess the risk of heteroskedasticity (unequal variance of residuals across the variable values). This would have had implications for the model and the tests used in the analysis. The variance of the residuals does not look constant over different fitted values. Subject to the SRS statistical disclosure guidance, we are not able to display the residual plots. We instead performed a Shapiro-Wilk test for normality check on residuals, which showed that the distribution of residuals departed significantly from normality (W = 0.99, p < 0.001). This indicates heteroskedasticity; hence, we used maximum likelihood estimation with heteroskedasticity-robust standard error in our analysis, despite a large sample.

Figure 9: Distribution of age-standardised GL PTM score band at endline by group for primary analysis—maths pupils eligible for FSM



a. Intervention group

b. Control group

In the multilevel model that accounts for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils) with randomisation strata and year group as a fixed effect, the adjusted difference in means is equal to 0.49 age-standardised PTM score, with a p-value of 0.423 (Table 14 and Table 15). The effect size associated with the adjusted difference in mean is 0.04 (95% CI: -0.06, 0.13), which suggests weak evidence of a very small positive effect size of the intervention on the GL PTM scores of FSM eligible pupils. The Hedges' g effect size of 0.04 translates to no additional month's progress.

The post-intervention intra-cluster correlations (ICCs) were estimated directly from the primary outcome measure from schools allocated to both conditions. The ICC for within schools is 0.05 (95% CI: 0.03, 0.09), obtained from the model with no adjustments.

Table 14: Primary outcome analysis results—maths attainment for FSM-eligible pupils

Outcome	**Unadjusted differences in means (I-C) (95% CI)**	**Adjusted differences in means (I-C) (95% CI)**	**Intervention group**		**Control group**		**Pooled variance**
			n (missing)	**variance of outcome**	**n (missing)**	**variance of outcome**	
Age-standardised PTM score	0.33 (-0.94, 1.60)	0.49 (-0.72, 1.70)	943 (272)	193.55	898 (262)	191.32	192.46

[37] The cut-off points are: (1) very low against the national average: < 74; (2) below average: 74–81, 82–88; (3) average: 89–96, 97–103, 104–111; (4) above average: 112–118, 119–126; and (5) very high: > 126.

Table 15: Primary outcome analysis—effect size estimation (maths attainment for FSM-eligible pupils)

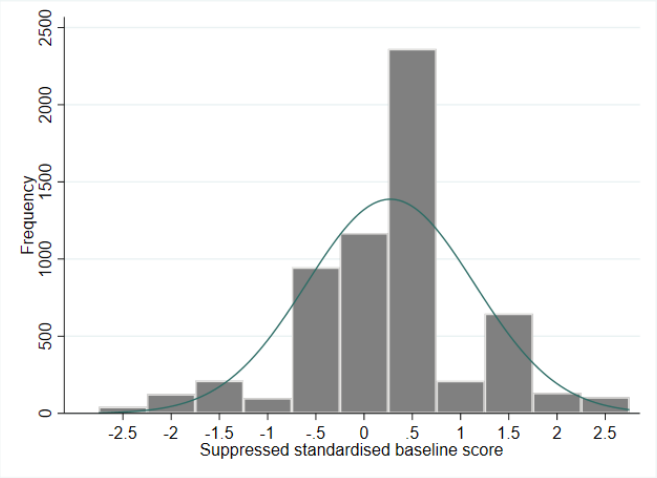
Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean 95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Age-standardised PTM score	943 (272)	89.64 (88.75, 90.53)	898 (262)	89.31 (88.41, 90.22)	1841 (943; 898)	0.04 (-0.06, 0.13)	0.423

Secondary analysis

Maths attainment for all Year 3 and Year 5 maths pupils

The analysis explored the impact of Stop and Think on maths attainment of all Year 3 and Year 5 pupils (RQ2). Pupils allocated to maths tests in the final analysis (6,044 pupils with observable data on all variables) had a mean of 0.07 for the standardised baseline measure (SD = 0.91; Skewness = -0.49; Kurtosis = 4.46). Here, again, subject to the SRS statistical disclosure guidance for presenting data, we grouped the standardised baseline measure for the analysis into eleven categories as for the primary analysis. Figure 10 presents the grouped standardised baseline score, with 39% (n = 2,361) of pupils obtaining a standardised baseline score greater than zero but below or equal to 0.5.

Figure 10: Distribution of grouped standardised baseline score—all maths pupils

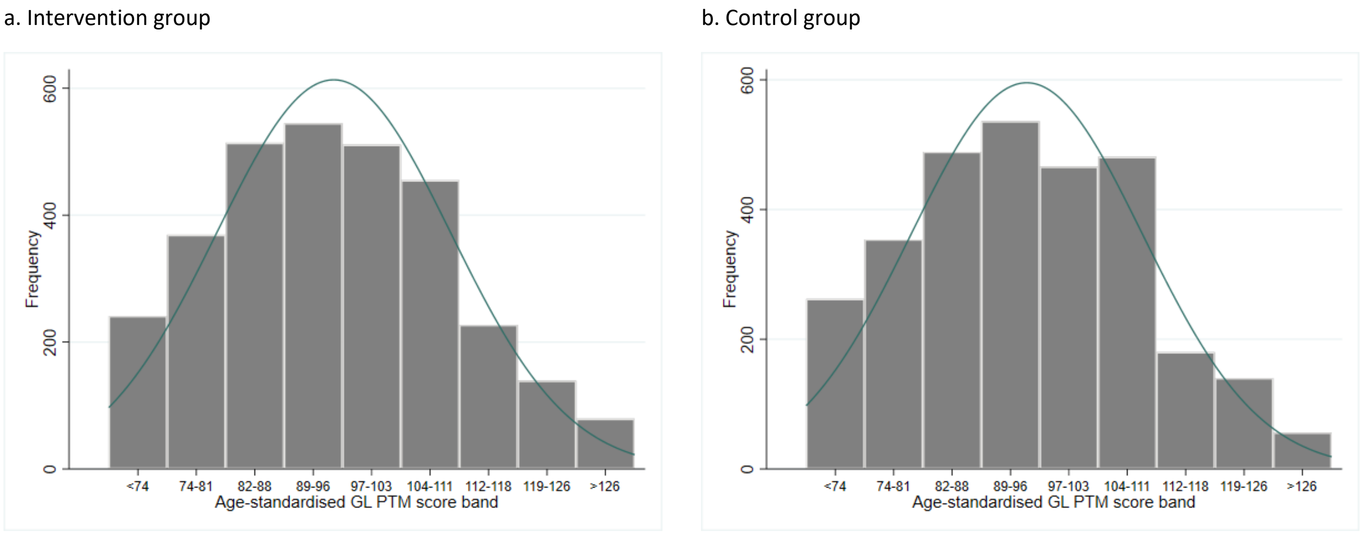


Pupils in this set of analysis had an overall mean of 94.80 and standard deviation of 15.17 for the age-standardised GL PTM at endline (Skewness = 0.31; Kurtosis = 2.66).[38] The overall mean of the maths test for these pupils were higher than pupils coming from a disadvantaged background identified by their eligibility for FSM. The intervention group had a mean of 95.11 and standard deviation of 15.29 for PTM at endline (Skewness = 0.33; Kurtosis = 2.69) while the control group had a mean of 94.48 and standard deviation of 15.04 (Skewness = 0.28; Kurtosis = 2.61).

The age-standardised GL PTM was moderately correlated to the standardised baseline measure with $r = 0.55$ ($p < 0.001$) among all maths pupils. To comply with the disclosure guidance, we present the age-standardised GL PTM score as nine categories as we did for the primary analysis. The distribution of the primary outcome—age-standardised GL PTM score bands—is shown in Figure 11. The figure shows a peak at the band of 89–96, with around 18% (n = 1,081) of pupils obtaining a score in this band. Here, again, the observations were not normally distributed and plotting the residuals against the fitted values shows sign of heteroskedasticity, hence we used robust standard errors in our analysis.

[38] Similar to the primary analysis sample, the statistics suggest that the distribution of age-standardised GL PTM at endline among all Year 3 and Year 5 maths pupils was symmetrical, though light-tailed.

Figure 11: Distribution of age-standardised GL PTM score band at endline by group—all maths pupils



Unadjusted differences in means of age-standardised PTM scores between the intervention and control group is 0.63 (Table 16). In the multilevel model that accounts for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils), randomisation strata and year group, the adjusted difference in means of age-standardised PTM score is equal to 0.66, with a p-value of 0.174 (see Table 16 and Table 17). The effect size associated with the adjusted difference in mean is 0.04 (95% CI: -0.01, 0.09), which suggests weak evidence of a very small positive effect size of Stop and Think on the GL PTM scores of all maths pupils. Similar to the primary analysis, the Hedges' g effect size translates to no additional month's progress.

Table 16: Secondary outcome analysis results—maths attainment for all pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Age-standardised PTM score	0.63 (-0.14, 1.39)	0.66 (-0.29, 1.61)	3080 (641)	233.72	2964 (609)	226.21	230.03

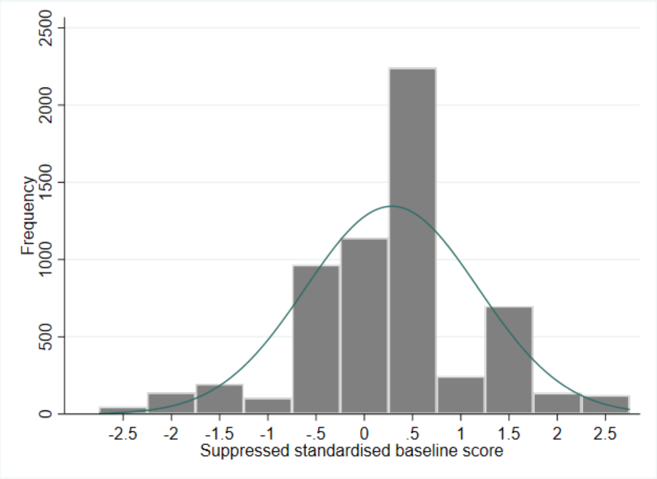
Table 17: Secondary outcome analysis—effect size estimations (maths attainment for all pupils)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Age-standardised PTM score	3080 (641)	95.11 (94.57, 95.65)	2964 (609)	94.48 (93.94, 95.02)	6044 (3080; 2964)	0.04 (-0.01, 0.09)	0.174

Science attainment for all Year 3 and Year 5 science pupils

The analysis explored the impact of Stop and Think on science attainment of all Year 3 and Year 5 pupils (RQ3). For pupils allocated to science tests in the final analysis (6,033 pupils with observable data on all variables), the standardised baseline measure had a mean of 0.08 (SD = 0.94; Skewness = -0.55; Kurtosis = 4.69). We also grouped the standardised baseline measure into eleven categories for display, as we did for the primary analysis to align with SRS statistical disclosure guidance. Figure 12 presents the grouped standardised baseline score, with 37% (n = 2,241) of pupils obtaining a standardised baseline score greater than zero but below or equal to 0.5.

Figure 12: Distribution of standardised baseline score—all science pupils



Pupils in this set of analysis had an overall mean of 102.46 and standard deviation of 17.48 for the age-standardised GL PTS at endline (Skewness = 0.16; Kurtosis = 2.57). The intervention group had a mean of 103.11 and SD of 17.69 for PTS at endline (Skewness = 0.15; Kurtosis = 2.56) while the control group had a mean of 101.78 and SD of 17.22 (Skewness = 0.16; Kurtosis = 2.56).

The age-standardised GL PTS was moderately correlated to the standardised baseline measure with *r* = 0.52 (*p* < 0.001) among all science pupils. Here, again, to comply with the SRS statistical disclosure guidance, we grouped the age-standardised GL PTM score into nine categories to present the distribution of the primary outcome—age-standardised GL PTM score bands in Figure 13. The figure shows a peak at the band of 104–111, with around 20% (n = 1,174) of pupils obtaining this score band and exhibits a non-normal distribution. The plotting of the residuals against the fitted values shows signs of heteroskedasticity; hence, robust standard errors were used in the analysis.

Figure 13: Distribution of age-standardised GL PTS score at endline by group—all science pupils

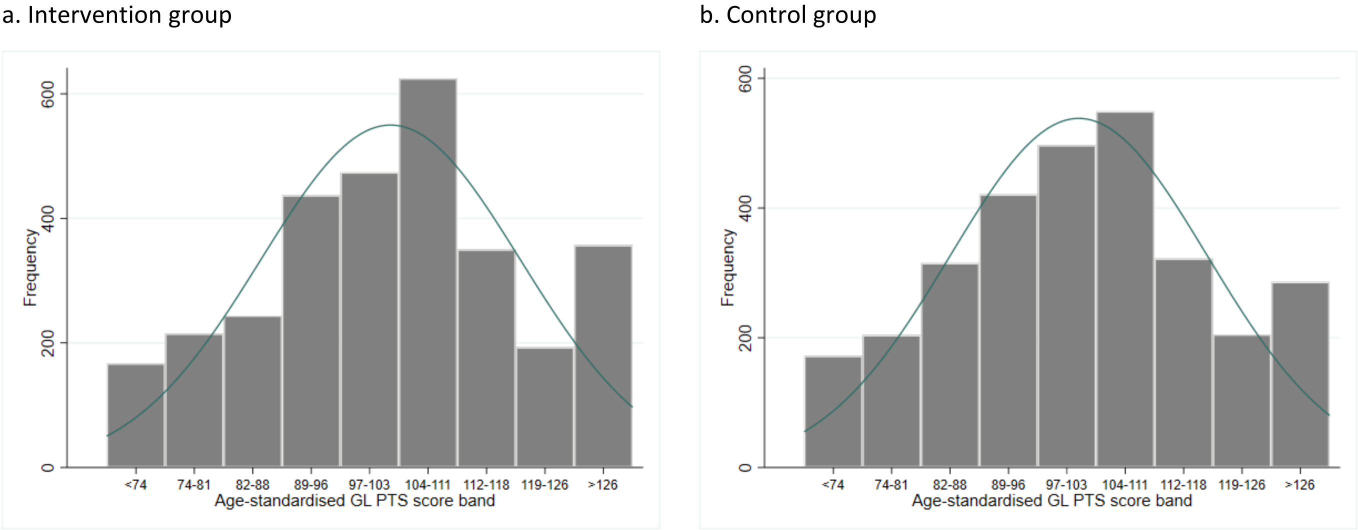


Table 18: Secondary outcome analysis results—science attainment for all pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	N (missing)	variance of outcome	
Age-standardised PTS score	1.33 (0.45, 2.21)	1.84 (0.76, 2.92)	3062 (656)	313.11	2971 (604)	296.63	304.99

Table 19: Secondary outcome analysis—effect size estimations (science attainment for all pupils)

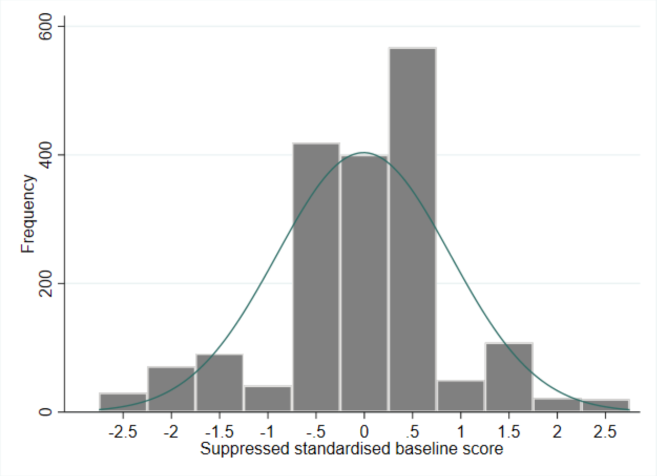
Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Age-standardised PTS score	3062 (656)	103.11 (102.49, 103.74)	2971 (604)	101.78 (101.16, 102.4)	6033 (3062; 2971)	0.11 (0.05, 0.16)	0.001

Unadjusted differences in the means of age-standardised PTS score between the intervention and control group is 1.33. After accounting for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils), randomisation strata, and year group in the multilevel model, the adjusted difference in means of age-standardised PTM score is equal to 1.84 (*p* = 0.001: see Table 18 and Table 19). The effect size associated with the adjusted difference in means is 0.11 (95% CI: 0.05, 0.16), which suggests evidence of a small positive effect size of Stop and Think on the GL PTS scores of all science pupils. The Hedges' g effect size of 0.11 translates to two additional months' progress in science, implying a larger positive effect size associated with the intervention on pupils' science attainment compared to that on maths.

Science attainment for science pupils eligible for FSM

The analysis explored the impact of Stop and Think on science attainment of Year 3 and Year 5 pupils from disadvantaged backgrounds, as measured by FSM status (RQ4). Pupils in the final analysis (1,819 pupils with observable data on all variables) had a lower mean of -0.23 than all science pupils for the standardised baseline measure (SD = 0.96; Skewness = -0.66; Kurtosis = 4.67). Figure 14 presents the grouped standardised baseline score by eleven groups, with 31% (n = 567) of pupils obtaining a standardised baseline score greater than zero but below or equal to 0.5.

Figure 14: Distribution of standardised baseline score—science pupils eligible for FSM



FSM-eligible pupils in this set of analysis also had a lower overall mean of 95.69 for PTS score at endline (SD = 15.96; Skewness = 0.34; Kurtosis = 2.83), as compared to all science pupils. The intervention group had a mean of 96 and standard deviation of 16.32 (Skewness = 0.36; Kurtosis = 2.88) for PTS at endline while the control group had a mean of 95.37 and standard deviation of 15.58 (Skewness = 0.31; Kurtosis = 2.74).

The age-standardised GL PTS was also moderately correlated to the standardised baseline measure with *r* = 0.47 (p < 0.001) among all FSM science pupils. Figure 15 illustrates the distribution of the age-standardised GL PTS score bands for FSM science pupils. The figure shows a peak at the band of 89–96, with around 19% (n = 337) of pupils obtaining this score band. Here, again, the observations are not normally distributed and plotting the residuals against the fitted values shows signs of heteroskedasticity, hence we used robust standard errors in our analysis.

Figure 15: Distribution of age-standardised GL PTS score at endline by group—science pupils eligible for FSM

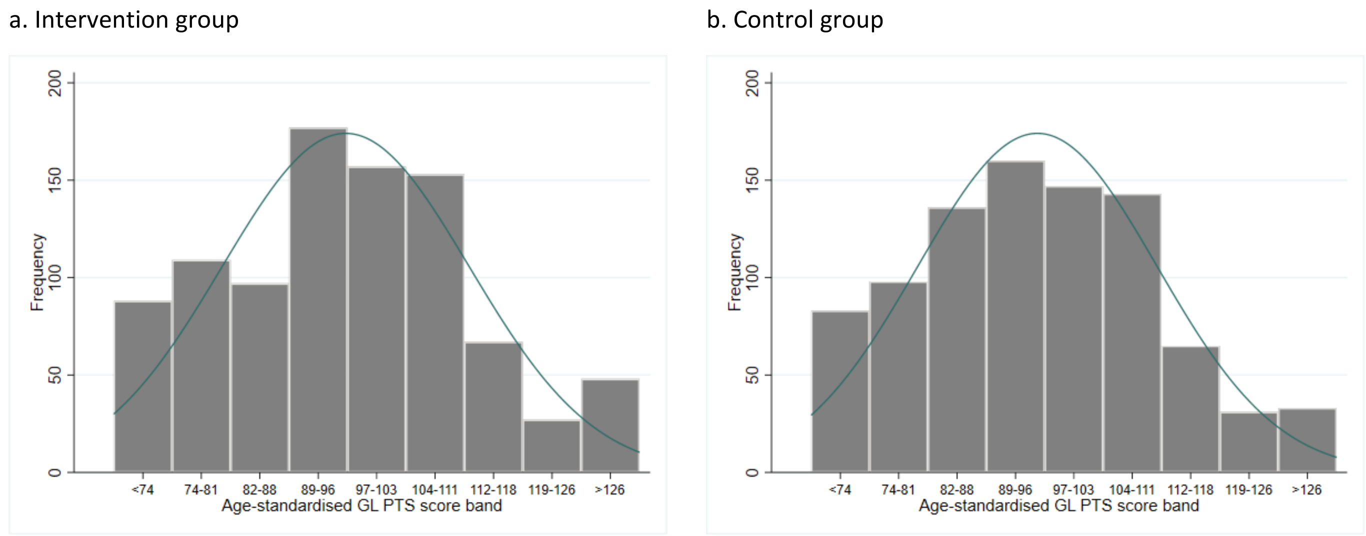


Table 20: Secondary outcome analysis results—science attainment for FSM eligible pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Age-standardised PTS score	0.63 (-0.83, 2.1)	1.25 (-.29, 2.8)	923 (270)	266.3	896 (263)	242.89	254.77

Table 21: Secondary outcome analysis – effect size estimations (science attainment for FSM eligible pupils)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Age-standardised PTS score	923 (270)	96 (94.95, 97.06)	896 (263)	95.37 (94.35, 96.39)	1819 (923; 896)	0.08 (-0.01, 0.17)	0.112

For FSM-eligible science pupils, unadjusted differences in means of age-standardised PTS score between the intervention and control group is 0.63. After accounting for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils), randomisation strata, and year group in the multilevel model, the adjusted difference in means of age-standardised PTM score is equal to 1.25 (p = 0.112: see Table 20 and Table 21). The Hedges' g effect size associated with the adjusted difference in means is 0.08 (95% CI: -0.01, 0.17). This effect size translates to one additional month's progress in science. However, the confidence interval includes negative values (ranges from -0.01 to 0.17) meaning that we are uncertain whether the intervention increases the GL PTS scores of FSM-eligible pupils.

Analysis for maths misconceptions for all Year 3 and Year 5 maths pupils

The analysis explored the impact of Stop and Think on all Year 3 and Year 5 pupils' misconceptions in maths (RQ5). The misconception test in maths consisted of five and nine validated items for Year 3 (α = 0.13; ω = 0.22) and Year 5 pupils (α = 0.48; ω = 0.33), respectively.[39] Note the reliability score did not reach an acceptable level by convention, which should be greater than 0.75 for a scale with less than 20 items (or 0.70 for a scale with more than 20 items, as suggested by Cortina, 1993). The technical report covering the development and validation of the misconception test

[39] The reliability assessment was based on the misconception scale (that is, misconception answer versus non-misconception answer) instead of subject scale (correct answer versus wrong answer). Given that the item responses were scored with a binary response scale, we used `R` functions within `lavaan` and `semTools` packages to obtain reliability estimate of alpha and omega, as suggested by Flora (2020).

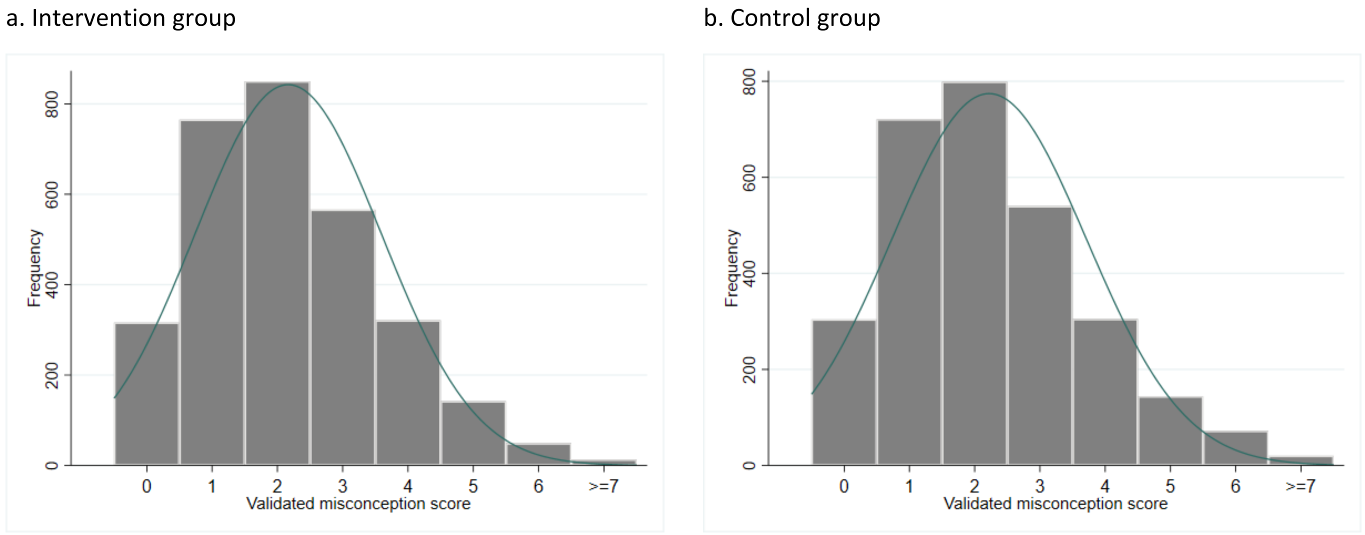
also indicates that the misconception score[40] achieved a reliability statistic (Cronbach's alpha) of 0.39 for Year 3 maths (round 1, 18 items, n = 345) and 0.43 for Year 5 maths (round 1, 22 items, n = 322). Consistent with the present evaluation report, the reliability score in the technical report did not meet the conventionally acceptable level either. Details on the validation and internal reliability statistics can be found in the technical report (McKaskill et al., in review).

Pupils allocated to the maths test in the final analysis (5,927 pupils with observable data on all variables) had an overall mean of 2.19 and standard deviation of 1.47 for the validated raw score of misconception (Skewness = 0.65; Kurtosis = 3.19).[41] The intervention group had a mean of 2.17 and standard deviation of 1.43 for the misconception score (Skewness = 0.62; Kurtosis = 3.17) while the control group had a mean of 2.22 and standard deviation of 1.50 (Skewness = 0.67; Kurtosis = 3.19). Here, again, following the SRS statistical disclosure guidance, we grouped the highest scores into the score of seven. The distribution of validated raw score of misconception is presented in Figure 16, with 620 pupils (around 11%) obtaining a score of zero and 33 pupils (less than 1%) obtaining a score of seven and above.

Unadjusted differences in means of raw misconception scores between the intervention and control group is -0.06. After we account for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils), randomisation strata, and year group in the multilevel model, the adjusted difference in means of raw misconception scores is equal to -0.07 (see Table 22 and Table 23).[42]

Table 23 further presents the effect size associated with the adjusted difference in mean which is -0.05 (95% CI: -0.10, 0.01). This suggests weak evidence that, on average, pupils in the intervention group were less likely to fall into misconceptions in maths. Note the 95% confidence intervals include zero, meaning that we are uncertain whether the intervention reduced pupils' misconception score in maths.

Figure 16: Distribution of validated raw score of misconception in maths at endline by group


a. Intervention group

b. Control group

Table 22: Secondary outcome analysis results—misconception in maths

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	

[40] The technical report for common misconceptions in KS2 maths and science coded the common misconception response as 'correct' to create a misconception scale.

[41] We calculated two measures of misconception: (a) missingness in a validated item was coded as 0 (that is, not falling into misconception) if a pupil was not absent from the misconception test and (b) missingness in any validated item was regarded as being absent from the misconception test. We carried out the misconception analysis for these two types of measures as an internal robust check. The results were consistent across two measures, so we only reported measure (a) for this report.

[42] There are five validated items for the Year 3 maths misconception test and nine for the Year 5 maths while there are seven items for Year 3 science and nine items for Year 5 science. Since we have already included year group in the model, this would adjust for the amount of opportunity a misconception event had. Hence, we did not include an offset term in the Poisson regression model to account that pupils' exposure to risk of falling into misconceptions which vary by test (year group or subject).

Raw misconception score	-.06 (-.13, .02)	-.07 (-.16, .03)	3022 (699)	2.05	2905 (668)	2.24	2.15

Table 23: Secondary outcome analysis—effect size estimations (misconception in maths)

Outcome	**Unadjusted means**				**Effect size**		
	Intervention group		**Control group**				
	n (missing)	**Mean (95% CI)**	**n (missing)**	**Mean (95% CI)**	**Total n (intervention; control)**	**Hedges g (95% CI)**	**p-value**
Raw misconception score	3022 (699)	2.17 (2.11, 2.22)	2905 (668)	2.22 (2.17, 2.28)	5927 (3022; 2905)	-0.05 (-0.10, 0.01)	0.157

Analysis for misconceptions in science for all Year 3 and Year 5 science pupils

The analysis explored the impact of Stop and Think on all Year 3 and Year 5 pupils' misconceptions in science (RQ6). The misconception test in science consisted of seven and nine validated items for Year 3 (α = 0.13; ω = 0.04) and Year 5 pupils (α = 0.19; ω = 0.12) respectively. Again, the reliability score was less than 0.75, which is the acceptable level by convention for a scale with less than 20 items (Cortina, 1993). Similarly to the present evaluation report, the technical report also indicates that the misconception score achieved a reliability statistic (Cronbach's alpha) of 0.06 for Year 3 science (round 1, 17 items, n = 344) and 0.16 for Year 5 science (round 1, 24 items, n = 324), which did not reach an acceptable level by convention either. Details on the validation and internal reliability statistics can be found in the technical report (McKaskill et al., in review).

Pupils allocated to the science test in the final analysis (5,978 pupils with observable data on all variables) had an overall mean of 3.18 and standard deviation of 1.53 for the raw score of validated misconception in science (Skewness: 0.23; Kurtosis: 2.73). The intervention group had a mean of 3.15 and standard deviation of 1.54 for the science misconception score (Skewness = 0.22; Kurtosis = 2.72) while the control group had a mean of 3.21 and standard deviation of 1.52 (Skewness = 0.25; Kurtosis = 2.74). Here, again, following the SRS statistical disclosure guidance, we grouped the highest score with the score of eight. The distribution of validated raw score of misconception is presented in Figure 17: 155 pupils (around 3%) scored zero and 11 pupils (less than 1%) scored eight and above.

Unadjusted differences in means of raw misconception score between the intervention and control group is -0.06. After accounting for standardised pre-trial test scores (KS1 maths outcome for Year 3 pupils and averaged EYFSP overall points score for Year 5 pupils), randomisation strata, and year group in the multilevel model, the adjusted difference in means of misconception score is equal to -0.07 (see Table 24).

Table 25 further presents the effect size associated with the adjusted difference in mean which is -0.05 (95% CI: -0.10, 0.00). This suggests weak evidence that, on average, pupils in the intervention group were less likely to fall into misconceptions in science. Note again the 95% confidence intervals include zero, meaning that we are uncertain whether the intervention reduced pupils' misconception score in science.

Figure 17: Distribution of validated raw score of misconception in science at endline by group

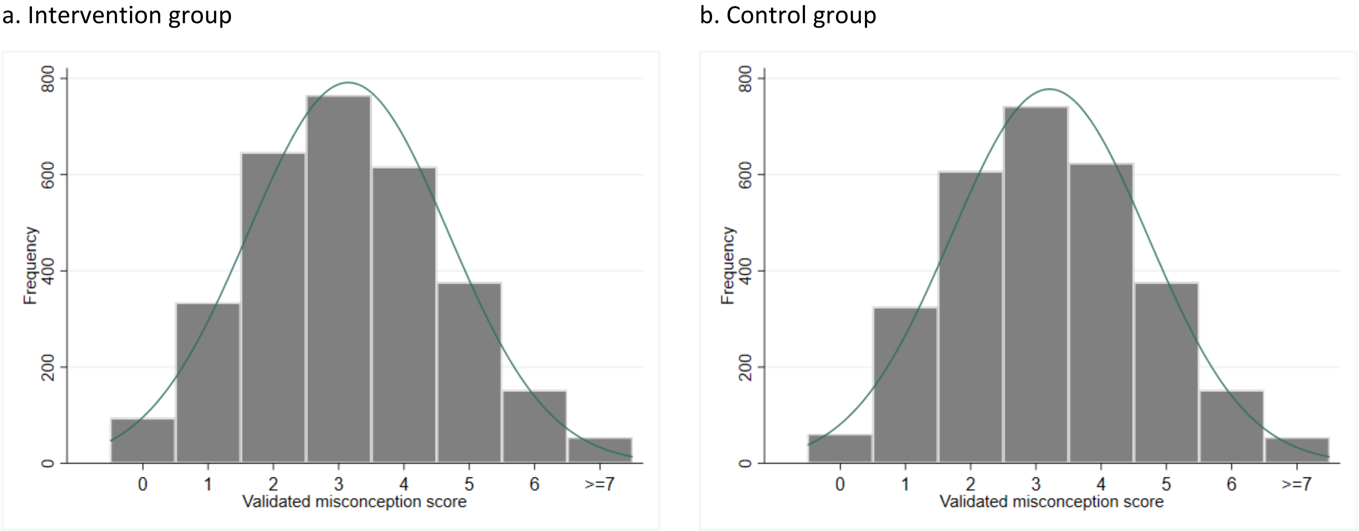


Table 24: Secondary outcome analysis results—misconception in science

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Raw misconception score	-.06 (-.14, .02)	-.07 (-.16, .01)	3037 (681)	2.36	2941 (634)	2.3	2.33

Table 25: Secondary outcome analysis—effect size estimations (misconception in science)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Raw misconception score	3037 (681)	3.15 (3.09, 3.2)	2941 (634)	3.21 (3.15, 3.26)	5978 (3037; 2941)	-0.05 (-0.10, 0.00)	0.078

Additional analyses and robustness checks

Sensitivity analysis for GL Assessment outcomes

As described in the Methods section earlier this report, we followed the approach used in the efficacy trial for the primary outcome analysis as well as other GL Assessment related outcomes. We undertook this analysis for comparability with the findings from the efficacy trial as well as complementing the IPE. For these sensitivity analyses, we ran two separate models for each year group and estimated the relevant effect sizes. Table 26 presents a summary of the results of the sensitivity analyses, while the full results can be found in Appendix H. For comparability, we also provide a comparison of effect size by year group and subject with the findings extracted from the efficacy trial (Roy et al., 2019) in Appendix H.

Our findings indicate a larger impact of the intervention on the GL PTM scores of all Year 3 pupils and the subgroup of Year 3 FSM-eligible pupils than Y5 pupils. However, there is some statistical uncertainty with both estimates. Furthermore, we combined the resulting effect sizes to estimate a single effect size across the two year groups. In line with our primary outcome analysis and secondary analysis of maths attainment for all pupils, the combined effect size for all and FSM-eligible pupils in maths was 0.04, which suggests weak evidence of a very small positive effect size of the intervention on the GL PTM scores of pupils. The Hedges' g effect size translates to no additional months' progress.

Our findings suggest strong evidence for a larger impact of the intervention on the GL PTS scores of all Year 3 pupils and Year 3 FSM-eligible pupils than Y5 pupils. The combined effect size across the two year groups for FSM pupils was 0.08 (95% CI: -0.01, 0.17), which translates to one additional month of progress in science. In line with the secondary analysis, this suggests weak evidence of a small positive effect size of Stop and Think on the GL PTS scores of FSM pupils. The confidence intervals marginally straddle zero and there is some statistical uncertainty with this combined estimate.

The combined effect size across the two year groups for all science pupils was 0.09 (95% CI: 0.04,0.13), suggesting evidence of a small positive effect size of the intervention on the GL PTS scores of all pupils. The combined effect size of 0.09 translates to one additional month of progress in science, which was slightly lower than that found in the secondary analysis of science attainment for all pupils (Hedges' g = 0.11; two additional months' progress).[43]

Overall, findings from the sensitivity analyses and the secondary analyses were consistent, both suggesting a larger positive effect of the intervention on pupils' science attainment compared to that on maths.

Table 26: Results of sensitivity analysis for maths and science attainment

Sample	Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	n in model (intervention; control)	Hedges g (95% CI)	p-value	Combined effect size Y3 and Y5 (95% CI)
FSM maths (primary analysis)	Raw PTM8 score – Year 3	0.59 (-0.91, 2.09)	0.82 (-0.67, 2.32)	899 (452; 447)	0.07 (-0.06, 0.2)	0.282	0.04 (-0.06, 0.12)
	Raw PTM10 score – Year 5	0.02 (-1.78, 1.83)	0.00 (-2.1, 2.1)	942 (491; 451)	0.00 (-0.13, 0.13)	0.998	
All maths	Raw PTM8 score – Year 3	0.28 (-0.58, 1.14)	0.95 (-0.23, 2.14)	3074 (1560; 1514)	0.08 (0.01, 0.15)	0.113	0.04 (-0.01, 0.09)
	Raw PTM10 score – Year 5	0.84 (-0.26, 1.93)	-0.03 (-1.85, 1.8)	2970 (1520; 1450)	0.00 (-0.07, 0.07)	0.976	
All science	Raw PTS8 score – Year 3	0.96 (0.46, 1.45)	1.15 (0.45, 1.85)	3049 (1524; 1525)	0.16 (0.09, 0.24)	0.001	0.09 (0.04, 0.13)
	Raw PTS10 score – Year 5	0.07 (-0.56, 0.69)	0.15 (-0.84, 1.13)	2984 (1538; 1446)	0.02 (-0.05, 0.09)	0.769	
FSM science	Raw PTS8 score – Year 3	0.92 (0.00, 1.84)	1.23 (0.22, 2.24)	872 (436; 436)	0.18 (0.04, .31)	0.017	0.08 (-0.01, 0.17)
	Raw PTS10 score – Year 5	-0.51 (-1.59, 0.58)	-0.06 (-1.32, 1.2)	947 (487; 460)	-0.01 (-0.13, 0.12)	0.929	

Subgroup analysis—FSM versus non-FSM pupils in the intervention group

We estimated alternative models to assess whether the findings for the primary analysis and science attainment for FSM-eligible pupils are robust to different model specifications. The alternative models explored the potential differential effect of the intervention on FSM-eligible pupils by including an interaction term between the treatment status and a dummy variable indicating FSM eligibility status. This additional analysis focused specifically on assessing whether the Stop and Think programme had a significantly higher effect on FSM-eligible pupils in the treatment group compared to

[43] Our original primary and secondary analyses pools Year 3 and Year 5 pupils together under the assumption of a common treatment effect, whereas the sensitivity analyses, which uses the meta-analytic approach, combines separate effect sizes, weighting them by their precision. This process accounts for potential heterogeneity in treatment effects across year groups and incorporates variability between them. As a result, the combined effect size might be slightly smaller and more conservative, reflecting this additional nuance.

their peers in the treatment group. Results for maths attainment are reported in Table 27 and Table 28, while those for science attainment are reported in Table 29 and Table 30. To obtain the effect size, the below difference in means was divided by the estimate of the population standard deviation in the estimation sample.

The Hedges' g effect size of -0.02 (95%CI: -0.09, 0.06) for both maths and science translates to no additional month's progress in either subject for FSM-eligible pupils in the treatment group compared to their peers in the treatment group. In other words, there is no evidence suggesting a statistically significant differential effect on maths or science attainment for FSM-eligible pupils receiving the intervention as compared to their non-FSM-eligible peers, given the confidence intervals covering zero.

Table 27: Subgroup analysis for maths attainment—FSM and non-FSM pupils in the intervention group

Outcome	Unadjusted differences in means (FSM-non-FSM) (95% CIs)	Adjusted differences in means (FSM-non-FSM) (95% CIs)	FSM group		non-FSM group		Pooled variance
			n (missing)	variance of outcome	N (missing)	Variance of outcome	
Age-standardised PTM score	-0.47 (-2.08, 1.15)	-0.25 (-1.62, 1.12)	943 (272)	193.55	2137 (354)	232.55	220.61

Table 28: Subgroup analysis for maths attainment—FSM and non-FSM pupils in the intervention group, effect size estimations

Outcome	Unadjusted means				Effect size		
	Intervention group: FSM		Intervention group: non-FSM				
	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total n (intervention)	Hedges g (95% CI)	p-value
Age-standardised PTM score	943 (272)	89.64 (88.75, 90.53)	2137 (354)	97.52 (96.87, 98.16)	3080	-0.02 (-0.09, 0.06)	0.720

Table 29: Subgroup analysis for science attainment – FSM and non-FSM pupils in the intervention group

Outcome	Unadjusted differences in means (FSM-non-FSM) (95% CIs)	Adjusted differences in means (FSM-non-FSM) (95% CIs)	FSM group		Non-FSM group		Pooled variance
			n (missing)	variance of outcome	n (missing)	Variance of outcome	
Age-standardised PTS score	-0.98 (-2.84, 0.88)	-0.30 (-1.95, 1.34)	923 (270)	266.3	2139 (368)	302.18	291.37

Table 30: Subgroup analysis for science attainment—FSM and non-FSM pupils in the intervention group, effect size estimations

Outcome	Unadjusted means				Effect size		
	Intervention group: FSM		Intervention group: non-FSM				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Effect size Total n (intervention)	Hedges g (95% CI)	p-value
Age-standardised PTS score	923 (270)	96 (94.95, 97.06)	2139 (368)	106.18 (105.45, 106.92)	3062	-0.02 (-0.09, 0.06)	0.719

Subgroup analysis: intervention versus control group in the FSM-eligible subsample

For comparability of the effect size between the interaction models and the main analyses of the FSM-eligible subsample (RQ1 and RQ4), we also extracted the effect size from the interaction models following EEF statistical analysis guidance (2022). For clarity, we summarise the results in Table 31, with the full results reported in Appendix H.

Although we used a different calculation approach, the effect size of 0.04 for maths pupils eligible for FSM in the interaction model is analogous to the primary analysis. In line with the main analyses, the subgroup analysis results suggest that the intervention did not have a statistically significant differential effect on maths or science attainment for FSM-eligible pupils given the confidence intervals covering zero for each raw interaction coefficient.

However, the effect size of 0.10, translating into two additional month's progress in science, is slightly larger than that found in the secondary analysis for science pupils eligible for FSM (ES = 0.08; one additional month's progress). The slightly larger effect size extracted from the interaction model might reflect the additional explanatory power of the interaction term (though not significant in predicting science attainment) and the specific variance considerations for the FSM subgroup.

Table 31: Subgroup analysis—results extracted from the interaction models

Outcome	Raw interaction coefficient (95% CI)	FSM subgroup effect size	p-value
Age-standardised PTM score	-0.25 (-1.62, 1.12)	0.04	0.720
Age-standardised PTS score	-0.30 (-1.95, 1.34)	0.10	0.719

Mediation analysis for maths pupils eligible for FSM

Out of the primary analysis sample—pupils assigned to sit maths tests and eligible for FSM—PTM score and maths misconception data was available for 1,863 pupils. Maths misconception scores were derived from misconception tests as the number of validated items for which pupils fell into a misconception.[44]

The two-level linear mixed effects regression model we estimated as step one of the causal mediation analysis indicated weak evidence of an only negligeable effect of the programme on maths misconception scores (b = -0.03, p = 0.612), adjusting for year group, baseline attainment and randomisation strata, and allowing the intercept to vary by school. This is in line with findings of our secondary analyses discussed above (see Misconception Analysis).

The two-level linear mixed effects regression model we estimated as step two of the causal mediation analysis further indicated weak and statistically uncertain evidence of an effect on PTM scores (b = 0.34, p = 0.525), accounting for maths misconceptions, year group, baseline attainment and strata, and allowing the intercept to vary by school. Maths misconception scores, however, were highly predictive of PTM scores across year groups and strata (b = -1.37, p < 0.001).

Hence, in line with our primary analysis, the causal mediation analysis showed weak and statistically uncertain evidence of a total effect of the programme on PTM scores among FSM pupils (0.38, 95% CI: -0.61; 1.47, p = 0.51). There was weak and statistically uncertain evidence of an average direct effect (ADE) on PTM scores (0.33, 95% CI: -0.64; 1.41, p = 0.55). There was also weak evidence of a negligeable average causal mediation effect (ACME: 0.05, 95% CI: -0.12; 0.23, p = 0.56). The proportion mediated effect was also negligeable at 0.06 (95% CI: -2.25; 1.51, p = 0.73).

Hence, there is evidence that maths misconceptions predict maths attainment, but we find weak and statistically uncertain evidence that the programme affects misconception scores or maths attainment, directly or indirectly.

[44] As discussed in the Secondary Outcome Analysis section above, the misconceptions test did not meet a conventionally acceptable level of reliability. Details on the validation and internal reliability statistics of the misconceptions test can be found in the technical report (McKaskill et al., in review).

Analysis in the presence of non-compliance

At randomisation, 87 schools were randomised into Condition 1 and 86 schools were into Condition 2. However, due to a mix-up caused by two schools having a similar name, pupils in one school allocated to the control group received the intervention. Our compliance analysis took this into consideration by allowing these pupils to have a non-zero number of completed sessions. A CACE analysis for the primary outcome was performed on complete cases only.

We carried out two sets of compliance analysis with two compliance measures: (a) compliance truncated to the intervention end date and (b) compliance untruncated until the endline testing. The results were consistent across two measures. For ease of interpretation, we only present results for measure (b) in Table 32. The first stage of the analysis that regressed the compliance indicator on the intervention status shows a coefficient on the latter of 23.79 (F (11,1785) = 1273.55; p < 0.001). The second stage of the analysis found an adjusted difference in means equal to 0.01 (p = 0.494) under the ITT analysis (0.49, Table 30). The associated effect size (Hedges' g) is equal to 0.00 (-0.09, 0.09), smaller than that observed in the ITT analysis (0.04, Table 15). This suggests that one cannot reject the null hypothesis that the intervention had no effect on the GL PTM scores of FSM eligible pupils, thus no evidence is found that increased compliance with Stop and Think leads to better maths attainment for FSM pupils.

Table 32: CACE analysis for the primary outcome

	Total n	Predictor	Adjusted difference in means	Effect size (95% CI)	p-value
IV model: stage 1—compliance indicator regressed on intervention status	1,797	Intervention status	23.79	N/A	<0.001
IV model: stage 2—PTM regressed on compliance indicator from stage 1	1,797	Compliance indicator	0.01	0.00 (-.09, .09)	0.494

Missing data analysis

There was some pupil- and school-level attrition from schools being notified of their randomisation allocation to the endline analysis. As shown in Table 11 in the Attrition section, the analysis included 12,077 of 14,718 pupils at randomisation of treatment allocation, equating to 17.94% pupil-level attrition (18.2% in the treatment group and 17.6% in the control group) and 3.47% school-level attrition (from 173 to 167 schools).

Following an ITT approach, the missing data analysis focuses on pupils for the primary outcome—maths pupils eligible for FSM—at the randomisation stage. However, where a pupil (or a school where pupils were nested in) withdrew from the study at the NPD consent stage, all data for that observation (including pseudo-identifiers, PTM, baseline scores, FSM status, and other covariates) became unavailable because we were not able to process or link their data in the SRS without consent. Consequently, these cases were excluded in the sample for the missing data analysis as they exhibited complete missingness. We instead examined the sample at the NPD consent stage as this represented the most available data accessible via the SRS.

The pupil-level attrition from the NPD consent to the analysis stage for the primary outcome is 22.48% (from 2,375 to 1,841 pupils), which is above the accepted level at 5% where patterns of missingness can be ignored. We thus ran a drop-out model to explore the pattern of missingness.

Following the SAP, our model to explore the pattern of missingness used a multilevel logistic regression to regress a binary variable indicating whether the primary outcome data (GL-PTM) was missing or not, and whether covariates in the primary model—treatment allocation, randomisation strata, year group, school identifier, and baseline attainment—were missing (or not). The covariates used to explore the pattern in missingness include treatment allocation, randomisation strata, year group, class-form entry size, school-level proportion of pupils eligible for FSM at any time during the previous six academic years, eligibility for FSM, imputed standardised baseline score, and whether baseline score was missing,[45] in addition to a random effect for schools. The analysis included all maths pupils at the NPD

[45] Standardised baseline score was replaced with the sample mean for missing observation in order to examine the pattern of missingness. Missing data for EVERFSM_6_P_SPR23 was coded up as the third category of FSM eligibility but not as a separate binary variable. In this way, missing data was retained in the model to explore the pattern of missingness. To clarify, the multiple imputation (MI) process did not involve any mean-imputed scores.

consent stage rather than those for the primary analysis (that is, FSM-eligible maths pupils only), taking into account the effect of missingness in the variable EVERFSM_6_P_SPR23 on identifying the primary sample. The coefficients of the drop-out model regressions are presented in Table 48 in Appendix F.

Table 48 suggests that the pattern of missingness was not related to treatment allocation but is statistically correlated with observable variables. Table 49 in Appendix G illustrates the probability of missingness in the outcome measure of PTM as a way to inform the multiple imputation (MI) process. We found that missingness in the primary outcome—age-standardised GL PTM—significantly correlated with pupils' FSM eligibility and baseline score. In fact, FSM-eligible pupils or pupils with a lower standardised baseline score were associated with a higher probability of missing the PTM outcome. None of the covariates apart from these two variables were significantly correlated with missing data. At the SAP stage, we had planned to use only significant variables to impute the primary outcome. However, we decided to include all variables used in the analytic model and auxiliary variables in the MI model to account for MAR and aid in building missing data models (see Methods section for details).[46] Here, again, the imputation process involved all maths pupils rather than merely those who were FSM-eligible.

The MI analysis was carried out using the `mi` Stata command, which uses a chained equation approach (MICE). Considering the multilevel structure of the data, we additionally included school-level clusters as dummy indicators in the MI model, accounting for the association between PTM outcomes and the partially observed variables within clusters. Based on the fraction of missing information, the minimum number of imputed datasets should be 20 as suggested by Graham at. al. (2007). However, 100 imputed datasets were generated in total, taking into account the convergence of the resulting parameters. Figure 26 in Appendix G illustrates convergence of the primary outcome (PTM), where it shows full levels of convergence.

Replicating the primary model, the PTM model for FSM-eligible pupils was then fitted across all imputed datasets. Table 33 and Table 34 report the re-estimated results using imputed datasets. The re-estimated results of the primary analysis through MICE remain consistent with the complete case analysis. The adjusted difference in means for age-standardised PTM between treatment and control FSM-eligible pupils amounts to 0.43, against 0.49 for the complete case (Table 14). A smaller effect size was also observed in imputed data, with the Hedges' g equal to 0.03 (CI: -0.05, 0.11), which is slightly smaller than that of the complete case analysis (0.04, CI: -0.06, 0.13). This suggests weak evidence of the Stop and Think intervention improving FSM-eligible pupils' maths attainment. The Hedges' g effect size translates to no additional month's progress.

Table 33: Analysis of the imputed datasets —maths attainment for FSM eligible pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Age-standardised PTM score	-0.14 (-1.42, 1.14)	0.43 (-0.72, 1.58)	1156 (59)	249.33	1114 (46)	238.35	243.94

Table 34: Analysis of the imputed datasets—effect size estimation (maths attainment for FSM eligible pupils)

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value
Age-standardised PTM score	1156 (59)	88.3 (87.46, 89.28)	1114 (46)	88.51 (87.60, 89.42)	2270 (1156; 1114)	0.03 (-0.05, 0.11)	0.467

[46] The variables used in the MI model included: treatment allocation, randomisation strata, year group, class-form entry size, school-level proportion of pupils eligible for FSM at any time during the past six academic years, eligibility for FSM, standardised baseline score, whether the baseline score was missing, and a factor variable indicating school-level clusters to account for clustering. To reiterate, we did not use any mean-imputed scores in the imputation model as this might cause biased estimates.

Implementation and process evaluation results

This section discusses findings from our IPE. We start with a short summary of the key findings before discussing each IPE dimension in detail.

Summary of IPE findings

- The different Stop and Think activities mostly took place as intended. Except for a few technical issues, the software was perceived as user-friendly and easy to navigate. This facilitated strong teacher engagement, enabling a smooth delivery of game sessions.
- Teachers used the programme flexibly while still adhering to the prescribed model. All respondents in the teacher post-intervention survey confirmed that sessions were completed as a whole-class activity.
- The scheduling of Stop and Think sessions deviated somewhat from the intended model. However, it is important to note here that the wording in the logic model—captured at the start of the evaluation—was slightly different to the intended delivery model as envisaged by the developers and subsequently delivered to schools by the delivery team. This led to a somewhat inaccurate framing of the delivery model in the IPE instruments: in our IPE, we found that 64% of teachers in the post-intervention survey said that they had delivered sessions at the start of maths and science lessons while 31% reported that they had varied the timing of sessions. In the qualitative research encounters, teachers cited practical reasons related to staffing and timetabling as having made scheduling of sessions difficult. However, in both interviews and surveys, teachers were asked about whether they had delivered the sessions at the start of maths and science lessons rather than at the start of maths and/or science lessons.
- The programme dose was mostly adhered to. In the post-intervention teacher survey, 72% stated that they had delivered three sessions every week. There were 30 sessions in total; the introductory session did not include a specific maths and science topic so is not included in the analysis. Data from the software showed that 53% of classes completed the other 29 sessions during the ten-week delivery period and that most classes completed at least 25 sessions. Only two classes of 303 completed no sessions.
- There was mixed evidence that pupils were responding less impulsively to maths and science questions. In focus groups, some pupils affirmed this; others did not. In the post-intervention survey, only around one in four teachers thought Stop and Think had been impactful in this regard.
- A key finding was the perceived difference in the quality of the maths and science content. Teachers found the maths content to be aimed at lower-ability pupils and expressed that it did not significantly add to their usual maths lessons. They also reported more inaccuracies in the content compared to science. Both teachers and pupils considered the quality and challenge level of the science items to be better.
- There was some evidence that differences between usual practice science and maths lessons and Stop and Think were modest. In the post-intervention survey, only about a third (31%) of teachers reported that Stop and Think was different to usual maths lessons; around half (55%) reported the same for science.

Fidelity

IPE RQ1 To what extent do the delivery partners and teachers deliver Stop and Think as intended?

Under the broad heading of Fidelity we explore the extent to which Stop and Think was delivered as intended. It includes:

- *quality*—how well the programme was delivered;
- *adherence*—the extent to which the delivery adhered to the intervention model;
- *adaptations* made to delivery; and
- *dosage*—how much of the intervention was delivered.

The section starts by describing adherence and adaptations made to the programme and then describes the dosage of the programme elements. Programme delivery and any changes made to it are discussed in the context of the Stop and Think guidance.

The intervention involved a 'light-touch' training session where BIT met with treatment class teachers, gave them each a handbook, demoed Stop and Think for them, set them up on the Stop and Think site, and answered their questions. This was followed by a ten-week programme delivery period where teachers delivered the programme in the classrooms. Fidelity has been explored both in terms of the training session as well as the delivery of the Stop and Think programme.

Quality

IPE RQ2 How well is Stop and Think delivered?

Quality was included in the IPE to understand how participants perceived delivery. Quality describes how well the programme is delivered unlike fidelity, which looks at whether elements of the intervention were delivered. Data on quality was collected in surveys, interviews, and focus groups with teachers and school leads as well as through in-class observations. This data, together with the findings on responsiveness, can help describe how and why activities from the logic model, when delivered in a particular way, can lead to outcomes.

In general, participants were positive about the overall quality of the training and support available for Stop and Think. Teachers offered very positive feedback for the training sessions from open text responses in the survey as well as interviews, describing the sessions as having been delivered 'expertly'. They appreciated that the sessions were easy to understand, informative, and useful. Teachers felt prepared after receiving the training session as this helped them gain an overall understanding of the programme and how to use it in their classes. Teachers said that the training was delivered in a way that allowed them enough time to process the information before the programme began.

Similarly, feedback in relation to the BIT trainers who conducted training sessions in schools was wholly positive. In open text survey responses, teachers stated that they found the trainers friendly, supportive, and engaging in their explanations and that the trainers had explained the programme well. Furthermore, in interviews, teachers said they found the training materials—especially the Stop and Think handbook—useful in familiarising themselves with the game.

Teachers' feedback corresponded well with the findings from the post-intervention interview with the delivery lead at BIT. It was mentioned that the in-person training sessions were received positively and that the sessions had allowed them to build a rapport with teachers and share ways in which teachers could reach out to them. This made the visits valuable to the set-up of the programme.

Additionally, teachers found BIT responsive when it came to their queries. According to the post-intervention teacher survey, 88% were satisfied with the technical support they had received during the programme. In interviews, teachers and school leads who accessed technical support, particularly by email, reported that they received replies within the same day and, occasionally, within a few minutes.

A contrasting view among school leads and teachers in interviews was that improvements could be made to the training. One drawback that teachers in the post-intervention survey reported on was that the training sessions did not explore all the elements of the game in detail. Teachers with this view felt that the training sessions should be longer so that more time can be spent exploring the game before they are asked to deliver it to classes.

The length of time between the training session and the start of the programme delivery was highlighted as a barrier to delivery. School leads perceived that there was a risk that teachers would forget how to use specific aspects of the game before actual delivery in classes began. They suggested that the training session should be held closer in time to actual delivery of the game. Also, teachers suggested organising a recap session before the intervention period began so that they could practise playing the game.

In terms of the perceived quality of the programme itself, outside the training and support, this was shaped by teachers' and pupils' engagement with the Stop and Think game. Facilitators and barriers to engagement are discussed in more depth under Responsiveness. For example, the user-friendliness of the materials contributed positively to teachers' perceptions of quality while technical issues and content errors affected it negatively.

Adherence and adaptation

IPE RQ3 How, why, and to what extent are changes made to Stop and Think?

This section describes the extent to which Stop and Think was delivered as intended and the adaptations made to delivery. It covers both teachers' adherence to the intended delivery of the programme—playing the Stop and Think game for ten weeks—and BIT's pre-delivery training sessions.

Teacher training sessions

According to our interview with the delivery lead at BIT, their primary role throughout the intervention was to support delivery in schools, including demonstrating how the Stop and Think programme worked. In the autumn term of 2022, four BIT research assistants, their managers, and the BIT project lead received a training session from Birkbeck on how the Stop and Think game works and how the school training was run in the efficacy trial. The BIT team then developed the school training for the present trial, which was to be delivered to all participating schools. The training of school staff was described as a core intervention activity in the logic model.

BIT described that the schools were divided into four groups based on their geographic spread. Each research assistant was allocated a group of schools to which they delivered in-person training prior to the start of the intervention. BIT reported that the research assistants had delivered in-person training sessions in 170 of the 173 schools. Additionally, BIT stated that they had shared catch-up materials, including an online video and a handbook, with those teachers who were unable to attend the in-person training. BIT also offered online training sessions for any new teachers who joined the schools later in the school year.

According to training logs prepared by BIT, training sessions were usually carried out with the Stop and Think school lead and the Year 3 or Year 5 teachers, depending on which year group was receiving the intervention. This matched what teachers discussed in the post-intervention survey and in interviews. BIT training logs suggested that, in some instances, IT leads, deputy headteachers, and teaching assistants had also attended the in-person training sessions.

The training sessions were 50 to 60 minutes long. In interviews, BIT said that the aim of the training sessions was to demonstrate how the game works and help provide useful resources to teachers so that they can deliver the game to the pupils. Each training session consisted of:

- introducing the game to the teachers;
- explaining how to set up and schedule the game in the classrooms;
- registering either Year 3 or Year 5 on the game;
- completing a practice run of the game with the teachers;
- explaining when automated reminder emails will be sent;
- helping them understand who to contact with queries; and
- and explaining when it should be delivered—the Stop and Think guidance suggested that teachers should complete three sessions per week over a course of ten weeks.

In interviews, teachers mentioned that they were also provided with their own copy of the Stop and Think handbook, online videos, and an FAQ document as a part of the training session. BIT training logs showed that they had sent out catch-up materials to any teachers who were absent from the in-person training sessions.

Teachers described BIT's approach to training sessions as a useful and positive experience. In the post-intervention survey, almost nine out of ten (88%, n = 60) teachers felt they had received adequate information during the training session to deliver Stop and Think. Moreover, teachers and school leads highlighted that they were given demonstrations on how to use the game and how to make adaptations to the game by using motivational elements or choosing different topics. Teachers mentioned that they received adequate training materials and found the Stop and Think handbook informative and resourceful (see Quality section).

Game delivery period

Adherence to the Stop and Think game delivery was assessed through interviews, focus groups, and surveys with teachers and school leads. These were compared with guidance in the Stop and Think handbook and against the intended activities described in the logic model. According to the logic model, Stop and Think is delivered to pupils at the start of maths and science lessons over a ten-week delivery period. Here, again, it is important to note that the delivery team's intended delivery model was in fact for the sessions to be delivered at the start of maths and/or science lessons. Each game session was to last up to 12 minutes and guidance suggested that it needed to be played three times every week (see Dosage).

Delivery at the start of lessons

The post-intervention survey data reported that of the 69 teachers who completed the survey, 64% had delivered the programme according to the logic model—that they had had played the Stop and Think game at the beginning of both maths and science lessons. In interviews and focus groups, teachers supported this view with examples. Teachers and school leads mentioned that they used the game as a 'starter' activity before maths and science lessons. In the lessons that we observed, teachers were playing the game at the start of maths or science lessons. Teachers' accounts of playing the game suggest that this had been advised to them in the handbook and the training sessions.

Interestingly, several teachers indicated they had replaced some of their usual lesson time to fit in the Stop and Think game at the start of their maths and science lessons. This was contrary to the expectation of the developer which, in an interview with us before delivery started, stated that it did not envision any classes missing usual lesson time due to the game since it was very closely linked to the science and maths curriculum.

However, teacher survey data highlighted that almost a third (31%, n = 21) of teachers had varied their delivery of the sessions from the description in the logic model: that they had not delivered sessions at the start of maths and science lessons as opposed to maths and/or science lessons. In interviews and focus groups, teachers similarly expressed that they had not always delivered the game at the beginning of maths and science lessons. Observation data suggests that, in some scenarios, schools delivered the game at the start of reading lessons instead. The main examples of partial adherence to the logic model reported in teacher and school lead interviews, teacher survey, and focus groups and observation data included:

- playing only during a maths lesson or only during a science lesson (which was in line with the delivery team's intended model but contrary to the wording of the logic model);
- playing the game at the start of a maths lesson and at the end of a science lesson, or vice versa;
- playing the game at the start of a reading or an English lesson; and
- playing the game at the end of maths and science lessons.

When asked whether they delivered sessions at the start of maths and science lessons, teachers and school leads expressed a range of views about how common it was for them to do so. This was reported in interviews with teachers and school leads, the teacher survey, and the teacher focus groups. Some teachers and school leads said they did this throughout the delivery period, others only in a handful of scenarios. The following reasons for this variation were given.

- Timetabling restrictions: teachers and school leads faced difficulties fitting the game sessions into their timetable, so they preferred playing the game at whatever time was most convenient given their other commitments.
- Longer lessons: in some instances, lessons overran, and this meant that teachers had to move the game to the end of the school day to fit it in.
- Infrequent science lessons: teachers noted that their science lessons were not as frequent as maths lessons, so they preferred playing Stop and Think only at the start of maths lessons.
- Pupil behaviour: teachers in some schools noted that the game made pupils very excited and active, and that they would at times struggle to calm them down for their usual lesson after the game. We also witnessed this in our lesson observations: pupils were excited during the game and would shout out their answers and talk to each other while playing it. In response to this, one view among teachers was that they preferred playing the game at the end of their lessons so that they could manage this better.
- Mixed year groups: a small proportion of schools that participated in the programme had mixed year groups, that is, a year group containing a mix of Year 3 and Year 4 pupils or Year 5 and Year 6 pupils. In focus groups, teachers with mixed year groups reported that scheduling the game before a reading lesson was easier since pupils could play the game, while the other group of pupils who were not involved

in the programme could read their books. In the post-intervention survey, 5% of the 69 survey respondents (that is, three people) indicated having played the game before reading lessons.

Such instances of non-adherence to delivering sessions at the start of science and maths lessons included playing the game after school registration periods, after lunch, at the end of the school day, or as a standalone lesson. There is not enough evidence from the surveys and interviews to suggest that these arrangements were permanent throughout the delivery period. However, it does imply that, in certain instances, schools were not adhering to playing the game at the start of maths and science lessons as was described in the logic model.

Other factors considered for game delivery

Nine out of ten teachers (93%, n = 64) in the post-intervention survey reported that they typically played the maths and science games together in the same lesson. This matched what the game is programmed to do. In interviews, we heard from some teachers choosing between either playing the games together or separately. This does not match with how the game is programmed, so teachers may have misremembered or misreported the gameplay.

There is no specific guidance in the Stop and Think handbook on whether the game had to be played at a specific time during the school day. Teachers in interviews and focus groups did not widely comment on their preference when it came to delivering the game either in the morning or the afternoon. The timing of the sessions was dependent on when maths and science lessons were scheduled in their school timetables. However, in some instances teachers preferred morning sessions (because it was a better fit for the morning lessons and teacher schedules) or afternoon sessions (because pupils in these schools were divided into sets during morning classes). In focus groups, teachers further explained that for such a set-based structure, it was sometimes only appropriate to play the game in the afternoons when all the class was together again.

Optional motivational elements as game 'adaptations'

According to the handbook, motivational elements, such as a Stop and Think leader board and tokens that can be spent in the game Avatar Shop, are referred to as 'game adaptations.' The handbook also stated that teachers can change the order of the maths and science topics that are covered in each weekly session and select which sessions to cover in which weeks. However, it is important to highlight that such 'game adaptations' did not impact adherence to the fidelity of the programme. These motivational elements were optional and were explored to help understand the different ways that teachers interacted with the game.

Teachers varied in their usage of the leaderboard and tokens. Based on the accounts of teachers and school leads in interviews and focus groups, they either used both the elements or one element more than the other or instead of the other. For example, one school did not use the leader board but they spent the tokens; another did the opposite. Teachers said that they used whichever motivational element the pupils enjoyed and engaged with more. They related this to accounts of some classes being more competitive than others and wanting to use the leader board more. A contrary reason for teachers not using the leader board was to avoid encouraging competitiveness among the pupils.

One viewpoint among several teachers in focus groups was that, compared to Year 5 pupils, those in Year 3 were more likely to engage with the tokens, which they can use to personalise the avatars. This means that the year group could have also had an impact on which motivational element was selected between the tokens and the leaderboard.

Another new game adaptation was changing the weekly topics that would be covered in the sessions. This was introduced based on feedback from the efficacy trial where teachers had expressed a preference for more control over the order of the content. In interviews and focus groups, teachers and school leads described occasionally changing the topics in the game. This was usually done to match game topics with lesson topics. Observation data showed an example where a teacher changed the game's science topic to match what was going to be taught in the science lesson after the game had ended.

In addition, one view among teachers in focus groups highlighted that they at times changed the order of the topics in the game to cover the topics that the pupils had already studied in the school year. They felt that if they did not change the order, the teachers would have to conduct a 'mini teach' with their class each time a new topic was introduced. (Here, it is worth noting again that the aim of the intervention is to improve pupils' inhibitory control skills, not for them to gain knowledge of science and maths content. It is, however, understandable that teachers viewed the programme through the lens of content and prior knowledge in pupils as this aligns with their day-to-day role.)

> *'So, there were a couple of them [topics] that I adjusted so we would have encountered it in class before they came across it in the programme just so they had a bit of prior knowledge, and it wasn't completely new to them' (school lead, post-intervention school lead Interview 3).*

In the post-intervention survey, teachers described collecting tokens (34%, n = 24) and using the leaderboard (30%, n = 21) as the more commonly used adaptation when compared to changing the weekly themes (17%, n = 12). This was similar to our findings from interviews and session observations.

It is also worth noting that 18% (n = 12) of respondents stated that they had never used any of the motivational elements described above. One school lead interviewee's view was that teachers preferred to prioritise giving pupils enough time to think about their answers over using the leaderboard and tokens. Similarly, several school leads described teachers wanting the game to be 'short and sharp'; this suggests that using these elements was perceived to increase the length of the session. Another view from school leads in interviews was that pupils were already motivated by the game's content and since the aim of the game was to create a culture where it is okay to make a mistake, teachers did not want to use the leaderboard and tokens. Lastly, there was a viewpoint expressed in the teacher focus groups that being able to play the game was a reward in itself and that they had therefore explained to the pupils that it was not necessary to use the added motivational elements.

How classes interacted with the game

According to the Stop and Think handbook, teachers could decide how their class interacts with the game. Changes in class interaction are referred to as 'non-game adaptations'. Similarly to the game adaptations, these non-game adaptations did not impact fidelity since they were optional elements of the programme. Examples of such non-game adaptations included pupils recording answers on whiteboards or paper independently and then adding the answers in the game, pupils working out answers with a partner, and the different ways in which pupils were selected to give an answer.

Our pre-intervention interview with the developer also emphasised flexibility in the way teachers and pupils interact with the game. To ensure fidelity, the only stipulation, according to the developer interview, was that teachers should not accept the first or the fastest answer from pupils. Interview and survey data suggested that teachers were aware of this and made an effort to not select the fastest response as the answer.

A range of approaches on how pupils interacted with the game were discussed as a part of the surveys, interviews, and focus groups with teachers. All the described approaches gave all pupils the opportunity to think about and discuss the answer or to have a say in the final answer, thereby keeping them adherent to what Stop and Think intended. Teachers mentioned the following approaches, used on their own or in combinations:

- giving pupils time to think individually and using whiteboards independently to record their answers;
- group and paired discussions between pupils;
- each pupil sharing with the class what they thought was the correct answer and reasons why;
- randomly selecting individual pupils to answer using lollipop sticks or a randomisation software; and
- pupils being given control of the game and calling out their classmates for answers while the teacher monitored the class.

There was variation in how the final answer was picked for the game, according to teacher surveys and interviews. In most cases, the most popular answer was selected; this was decided by a show of hands or another type of class vote (for example, pupils holding their thumbs up or down or using red and green cards to indicate whether an answer was correct).

In our lesson observations, we saw a variety of ways in which classes were set up to interact with game:

- pupils stay at their desks and face the interactive board;
- pupils sit in a group on the floor and face the interactive board;
- some pupils sit on the floor and others remain in their chairs, which were arranged in a semi-circle;
- pupils sit in rows facing the interactive board.

Furthermore, teachers highlighted that they used tailored approaches for pupils with Special Educational Needs and Disabilities (SEND) and pupils with English as an Additional Language (EAL). School leads described the use of hand signals to support EAL pupils' understanding of the game. A less common approach involved making EAL pupils sit together so that they could help each other understand the game.

Framing questions differently for SEND pupils was also part of the flexibility that the game afforded teachers. School leads, in interviews, reported that questions which included word problems posed a challenge for SEND pupils and so re-framing them was useful. Also, in our lesson observations we noted that teachers would ask SEND pupils to answer at least one question every session so that they would not feel left out and could participate.

Communication and support provided during delivery

A number of communication and support channels were available to schools to support game delivery. Although schools and teachers were encouraged to use these resources, there was no obligation to do so.

Teachers were aware of the different support channels available to them, especially in scenarios where they faced technical difficulties with the game. BIT provided support to teachers who experienced technical issues with the game and teachers were asked to report any such issues to them, either via email or phone calls. According to the post-intervention survey, 22% (n = 15) of teachers sought support from BIT regarding technical issues during the duration of game delivery. Email was the most common way of seeking support. Teachers in the survey reported using both email as well as telephone contact with BIT. Teachers were positive about the interactions that they had with the BIT support team, as described in the Quality section.

We examined BIT's implementation logs for this evaluation. Alongside the results from open text responses in our post-intervention survey, they provide a record of the technical issues reported by teachers during the intervention. The more common included instances where the game was glitching, freezing, or taking a long time to load. Based on feedback from the delivery team, these issues were likely to have been caused by local broadband issues rather than something inherent in the software Other technical issues included:

- incorrect login details;
- issues with the game's audio;
- missing weekly badges;[47] and
- answers being perceived as marked incorrectly.

However, a much larger proportion of teachers (74%, n = 51) reported that they did not access any technical support during the game delivery period. In interviews, teachers and school leads said that they did not require additional support to deliver the game, giving a number of reasons for this. One view expressed that the game was easy to use and the support that they received in the training session had been adequate. Additionally, they stated that the intervention website was resourceful in helping them solve issues. Teachers were also able to fix some issues with buffering and enlarging graphics on their own. Buffering of images had been a major technical issue in the previous trial (NFER, 2019),[48] however, Birkbeck had made improvements for this new version of the game, which could explain why this did not come up as a technical issue.

Schools were proactive in communicating with BIT about the game for reasons other than requesting technical support. According to BIT implementation logs, popular reasons for contact included teachers reporting that they were unable to complete their three weekly sessions, either due to teacher strike days or staff absences. Also, schools contacted BIT to provide updates on staff changes which involved new teachers taking over game delivery. Another reason was teachers requesting updates on the number of sessions left to complete and querying why the game was no longer accessible (after they had completed all of the sessions).

[47] These badges were missing because the software did not account for school holidays (for example, February half term). This issue was not fixed during the trial. Instead, with the EEF's approval, the delivery team messaged all schools about the missing weekly badges just before the Easter holidays (and also pointed out that they would receive automated Friday reminders to play Stop and Think during the holidays, and that they could ignore these).

[48] NFER (2019) 'Stop and Think: Learning Counterintuitive Concepts: Evaluation Report' (pdf), Education Endowment Foundation: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Stop_and_Think.pdf?v=1716470273.

BIT's implementation logs showed that teachers provided feedback and reported issues like outdated academic language, difficult subject questions, and gave suggestions on how to improve game delivery. These issues were also raised by teachers in their interviews and are discussed in more detail in the Responsiveness section.

Finally, during the game delivery period, teachers received weekly automated emails every Friday, which prompted them to complete any missing sessions. If, by the Friday morning, a class had only completed one session so far that week, the reminder suggested that they complete one session that day, and four sessions instead of three the following week. Teachers in interviews confirmed that they had received these weekly emails and had found them helpful in planning game delivery.

Dosage

IPE RQ4 Do teachers deliver the intended dose of three 12-minute sessions for ten weeks?

This section explores how much of the intervention was delivered and whether this was in line with the intended dosage. Data on dosage was collected from interviews with teachers and school leads, teacher focus groups, the post-intervention survey, and software data collected by Birkbeck.

Three sessions per week

Teachers largely delivered the intervention three times per week. Of the 69 post-intervention survey respondents, 72% stated that they had delivered three sessions to their class every week, as was intended. This was also reported in interviews and focus groups by teachers. Furthermore, in interviews, school leads highlighted that the programme's structure of three sessions per week was 'doable' with no unintended consequences to the school curriculum.

In contrast, a little more than a quarter of the teachers (27%, n = 19) reported in the post-intervention survey that they 'sometimes' delivered three sessions per week. Instead, they reported that they delivered either two or four sessions in some weeks. According to the developer interview, in the initial efficacy trial of the intervention, almost half of the participating schools did not deliver the intended dose of three sessions every week.[49]

The reasons given by teachers in interviews and focus groups for being unable to complete the required number of sessions per week ranged from internal factors (such as staff strikes, inset days, and school trips) to external factors (for example, school closures due to snow). In particular, timetabling issues were cited as a cause of disruptions. Teachers explained that fitting in the three Stop and Think sessions into the same week was not always possible.

Ten-week delivery period

Participating schools had to complete a total of 30 sessions during the ten-week intervention period to complete the programme (note that 'weeks' here refer to school weeks). We have analysed the software data we received from Birkbeck to understand whether schools adhered to the intended delivery model. A few things are worth noting about the software analysis. First, the very first session was introductory and did not include a specific maths or science topic so has not been included in the analysis. Second, schools' access to the intervention was not cut off after ten weeks and they could continue playing any remaining sessions of the game after this point. Third, the intervention period was extended by one week. This was an extension offered by the EEF to all schools because of planned teacher strikes.

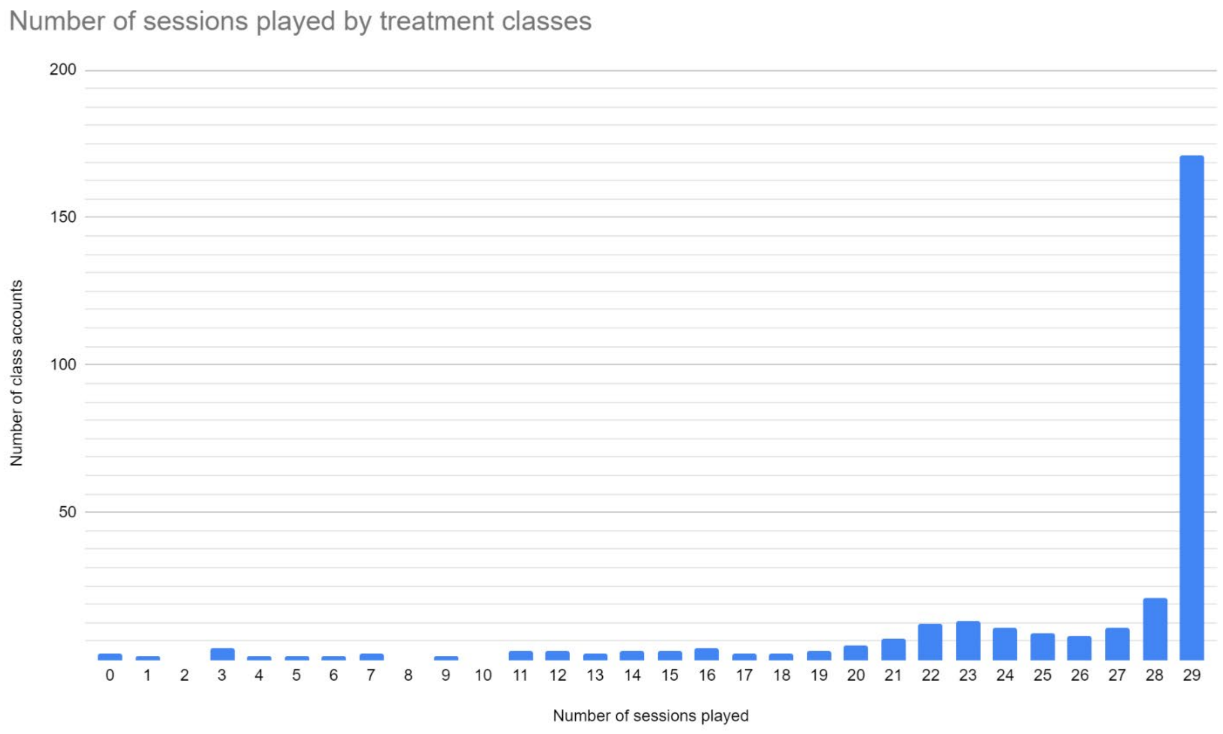
Our software data analysis found that:

- over half the classes (53%) completed all 29 sessions by the end of the ten-week delivery period;
- most classes completed at least 25 sessions after ten weeks;
- 22 classes (7%) continued to deliver sessions during this extension period and carried on until endline assessments;
- 15 schools (5%) finished the game within eight to nine weeks; we have no further information on why the programme ended early in these schools; and
- only two classes out of 303 did not complete any sessions.

Figure 19 shows the number of sessions completed by classes.

[49] We cannot comment on the exact number of sessions per week as we do not have the software data from the efficacy trial.

Figure 19: Number of sessions started at class level



In the post-intervention survey, 90% (n = 62) of teacher respondents said they had completed delivery in ten weeks. A small proportion reported that they delivered sessions for 11 to 15 weeks. Here, it is important to note that while the survey asked about delivery happening in ten weeks: the intervention delivery period was actually eleven weeks because of the extension granted by the EEF due to planned teacher strikes. Another small proportion mentioned that they had played the game for eight to nine weeks, indicating that they had covered the sessions faster than was intended.

In interviews, teachers and school leads discussed the reasons for not carrying out delivery over a ten-week period. The reasons were similar to those that were stated for not delivering three sessions every week. One reason for missing sessions was school events like trips; another was limited capacity due to staffing issues, including staff sickness. In some instances, school leads mentioned that BIT had extended the intervention period. This flexibility was received positively by school leads who stated that the extra time eased the pressure on teachers and allowed them to deliver the intervention as it was intended and without rushing.

> *'I think it just takes the pressure off teachers. It means it's done properly, and it's not rushed. You're not trying to cram it into where it doesn't naturally fit into the timetable, and I think that's ... You're going to get the best results when it's like that' (school lead, post-intervention school lead interview 4).*

Twelve-minute sessions

Each game session was programmed to last twelve minutes. This meant that there was no software data that could indicate whether it was played for longer than that time. However, in interviews, teacher and school leads reported that game play session length varied between twelve minutes (as intended) and those that played it for shorter or longer.

Teachers in interviews that had reported that the game lasted exactly twelve minutes stated that this was because the game automatically ended after that point. Others highlighted that bonus rounds and repeated questions could make the game seem like it lasted longer, even though it automatically ended after twelve minutes.

Another view among teachers and school leads, in interviews, was that it was common for sessions to last between 15 to 20 minutes. The main reason for this was the class discussions that took place during or after the game—the latter being evident in our observed lessons. However, it is important to point out that based on these accounts, teachers seemed to consider the time taken for discussions as part of the time it took to play the game when estimating timings.

Related to playing longer sessions, teachers experienced issues fitting in the sessions due to the time it took to deliver the game. For instance, a 45-minute maths lesson had to be replaced with the Stop and Think game as sessions involved many discussions and became too long to use as a starter activity.

> *'Owing to the time limitations in school, we've had to substitute a maths revision lesson for two of the Stop and Think sessions, as opposed to it being a precursor or a warm-up activity' (intervention teacher, school visit 3).*

Another explanation for sessions lasting longer was the time needed to set up: as discussed in the section on Adherence and Adaptation, in some schools, pupils used whiteboards and other resources to interact with the game, which required additional time for teachers to set up. It could be that teachers considered these set-up activities as a part of the session when they reported on timings. It was reported that as the intervention progressed, the classes became more adept and quicker at setting the game up.

There were rare instances when the game was reported as lasting for less than twelve minutes. In interviews, a view among teachers was that the game was 'short' and lasted only nine minutes. Reasons for this were discussed and fell into two broad categories: it was either due to technical issues like freezing, crashing, and being unable to select answer options or due to the pupils finding the questions too challenging, which made teachers end the game early to aid further discussions instead.

Reach

IPE RQ5 What is the rate and scope of participation at a school, class, and pupil level?

This section explores the rate and scope of participation in the Stop and Think intervention at the school, class, and pupil level. Compliance at the school level is not discussed here but under the Impact Evaluation Results section.

Initially, 173 schools were signed up to deliver the programme. The delivery team at BIT reported that 170 received an in-person training session covering 382 staff members in total. These sessions were completed at each participating school as it was a mandatory requirement for set-up of the game. According to BIT's training logs, 32 teachers did not attend the in-person training session. Reasons for absence included staff sickness, staff leave, a lack of lesson cover, or teachers having to attend other training.

To make up for any absences, BIT sent catch-up materials to absent teachers that included the Stop and Think handbook, online videos, and an FAQ document. Additionally, in the event of any staff changes, they sought to share videos and catch-up materials and to offer virtual training sessions to all new intervention teachers. However, school leads noted in interviews that not all new teachers received training or materials. The BIT training logs show that in some instances they were unable to convince new teachers to attend a virtual training session or they did not receive contact details for new teachers and so were unable to share resources or offer training directly.

Training sessions were usually carried out with the Stop and Think lead and the Year 3 or Year 5 teachers. The training logs show that other staff members such as IT leads, deputy headteachers, teaching assistants, and project leads also attended the in-person sessions on occasion.

One hundred and seventy schools participated in the endline testing in summer term 2023 as three schools had withdrawn from the programme at different points (see the Impact Evaluation Results section for more information). These schools withdrew before the intervention began so they did not complete any game sessions. The reasons for withdrawal were to do with staffing and other internal issues at the schools.

In interviews and focus groups, teachers and school leads at participating schools confirmed their intention to complete all 30 of the game sessions. They highlighted rescheduling specific sessions (see the Dosage section for more information) to ensure that the game was completed at their schools. Software data was collected from 301 individual classes (some schools had more than one class). The software data showed that 165 individual classes had delivered the entire programme at their respective schools within the ten-week period; an additional 70 classes had delivered at least 26 sessions. Additionally, schools were given an extension of one week by the EEF beyond the initial ten-week delivery period to complete all their pending game sessions and make up for any time that was lost due to teacher strikes.

Stop and Think was intended to be delivered as a whole-class activity. The interview with developers at Birkbeck highlighted that this was decided based on the pilot development phase.[50] They had found that many schools were unlikely to have the technological provisions available for pupils to play the game individually, which meant that a whole-class approach was preferred.

Data on pupils' participation in the intervention at an individual level was not collected as this would have placed significant evaluation burden on classroom teachers. However, the surveys, interviews, focus group, and observations all collected data on whether the game was delivered to all the pupils in the class or not. In the post-intervention survey, all (100%, n = 69) teachers reported delivering the programme as a whole-class activity. School leads and teachers supported this survey finding in their interviews and focus groups.

Although the game was typically played as a whole-class activity, school leads did report situations where some pupils were absent from lessons where Stop and Think was played. One reason was pupil involvement in another intervention or activity. Additionally, we observed during a school visit that a small group of pupils left the lesson at the beginning and the game was delivered as a whole-class activity to the remaining pupils. Reasons for leaving the lesson were not known to the researcher.

Responsiveness

IPE RQ6 How well do teachers and pupils engage with the intervention?

Teacher engagement

In this section, we discuss school lead motivations for signing up to the programme. This is followed by facilitators and challenges to teacher engagement.

Motivations for signing up

In endline interviews, school leads discussed motivations for signing up to the programme. Two main reasons were given: (1) the programme ethos and focus on misconceptions and (2) the anticipated ease of implementation of the programme.

For school leads, exposing the logic behind why misconceptions were incorrect was a key motivation for signing up. Another connected reason that school leads highlighted during interviews was a perception the programme would be accessible and inclusive for all pupils, including pupils with low socio-economic status and pupils with SEND. Teachers explained that these groups of pupils are particularly prone to misconceptions. Lastly, school leads mentioned a lack of science teaching interventions available to primary schools making Stop and Think particularly appealing.[51]

In cases where school leads discussed the expected ease of implementation, Stop and Think was described as being a 'pick up and go programme' that would not be overly burdensome for schools. Other reasons mentioned by school leads for signing up were:

- the programme aligning with current teaching strategies and practices;
- school commitment to 'evidence-based practice'; and
- school commitment to 'pedagogical improvement'.

Teachers were asked in the baseline teacher survey if they anticipated any challenges taking part in the trial: half (50%, n = 78) did not anticipate any challenges, 22% (n = 34) did anticipate challenges, and the remaining 28% (n = 44) were 'unsure' whether they anticipated challenges.

[50] The comparison between whole group and individual performance is published in: Wilkinson, H. R. , Smid, C., Morris, S., Farran, E. K., Dumontheil, I., Mayer, S., Tolmie, A., Bell, D., Porayska-Pomsta, K., Holmes, W., Mareschal, D., Thomas, M. S. C. & the UnLocke Team (2019) Domain-specific inhibitory control training to improve children’s learning of counterintuitive concepts in mathematics and science. Journal of Cognitive Enhancement. doi.org/10.1007/s41465-019-00161-4

[51] The lack of available science interventions was also mentioned in pupil focus groups: pupils commented that while they had previously used maths computer games at school, they had not used comparable science computer programmes.

Table 35: Whether teachers anticipate any challenges that could arise from taking part in the Stop and Think trial (baseline teacher survey)

	Frequency	Percentage
Yes	34	22%
No	78	50%
Unsure	43	28%
Total	155	100%

Of the 22% (n = 34) of teachers who anticipated challenges to participation, the most stated challenge was time constraints—that the programme would reduce time for curriculum teaching. Other potential challenges stated were:

- engagement and accessibility for all pupils, such as SEND and EAL pupils;
- concerns about the quality of the technology and/or whether it would be fit for purpose; and
- ease of usage of the technology.

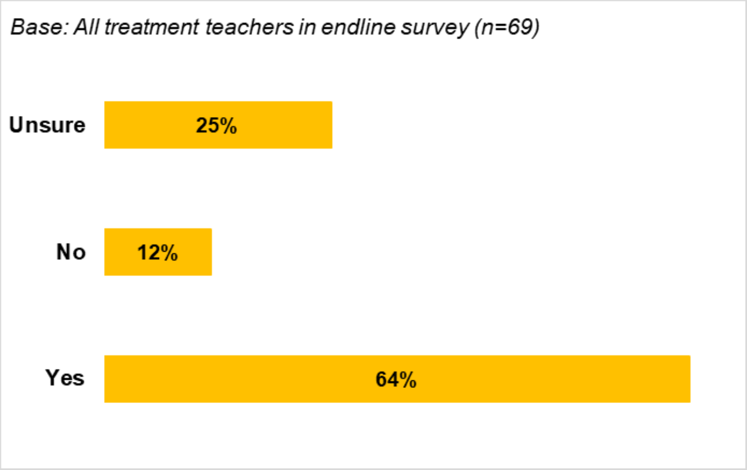
Facilitators to teacher engagement

In the endline teacher survey, over three quarters of teachers reported their experience of using the programme as either good (51%, n = 35) or very good (26%, n = 18). Over half (55%, n = 38) stated they would use it in the future. Teachers explained their responses further in open text questions. The most common reasons for high ratings were:

- high pupil enjoyment and engagement playing the game;
- focus on addressing misconceptions being beneficial to academic development; and
- ease of programme use.

We also asked teachers in the endline survey whether they would recommend the Stop and Think game to others: 64% said they would, 12% reported they would not, and the remaining quarter (24%) of teachers were unsure.

Figure 20: Whether teachers would recommend Stop and Think to others



We asked treatment teachers in focus groups, and school leads in interviews, what factors supported teacher engagement. A range of views were discussed, which resonated with the factors that motivated school leads to sign up for the programme initially (as detailed above).

First, teachers and school leads described how the programme ethos and focus on misconceptions corresponded with their teaching values and inclusivity objectives. Teachers commented that pupils were aware their peers may have different needs and abilities and the programme design supporting this understanding. Teachers complimented the programme as normalising making mistakes and getting things wrong, promoting this as a key part of learning.

> *'One of the things I really like about it is that it's all-around typical misconceptions and we celebrated making mistakes a lot … making mistakes I think is so, so important. Children having that resilience to feel that, great, this is an opportunity to learn something' (school lead 1, post-intervention school lead interview).*

Interestingly, pupils too commented on the primary aim of the programme being about promoting the quality of thinking:

> *'I feel like Stop and Think isn't important about getting it right, I feel like it's more to so, like it's better to learn' (pupil, pupil focus group 4).*

Second, teachers discussed a number of factors that made the programme easy to implement. They mentioned the quality and timing of the training and responsiveness of delivery support. Teachers who discussed this remarked favourably on communication being clear and support being 'readily available'. Training was described as easy to enrol on, unproblematic to access, and as having high quality accompanying materials.

The second factor was programme technology being perceived as user-friendly and easy to navigate, with equipment straightforward to use. In class observations we saw teachers confidently using the technology, not facing any technical issues, and using elements such as the leader board. In an interview that was conducted during programme delivery, one teacher mentioned that colleagues had not initially been confident with the technology, and they had spent time training them. This training resolved the issue, enabling colleagues to engage with the programme. This suggests that less-confident teachers were relatively easily able to learn how to use the technology and go on to confidently use it.

The third aspect was the flexible elements of the programme that gave teachers autonomy to use the programme in ways that best supported curriculum teaching. Teachers who highlighted this commented that it was useful to control when to revisit, and/or how to structure the units, so that they complemented usual lessons. For example, they were able to schedule subjects as a way of checking pupils' baseline knowledge or return to subjects to check on knowledge retrieval. During a classroom observation we saw this in practice, observing the teacher picking a theme about food chains as the following science lesson was linked to this topic.

Barriers to teacher engagement

Around 19% (n = 13) of teachers rated their experience as 'neither good nor bad'. Almost 17% (n = 12) said they would not use the programme again and 28% (n = 19) were unsure whether they would. The open text responses given for these survey questions tallied with the reasons respondents gave in qualitative encounters, as explained below.

In interviews and focus groups, teachers highlighted two main barriers to their engagement with Stop and Think: staff absences and timetabling. It is worth noting that these were similar to the reasons that teachers gave for missing their weekly sessions or not being able to complete the ten-week delivery period. In other words, the practical reasons that made the programme difficult to implement in schools were also barriers to teachers' engagement. Other barriers were to do with the content, technical issues, and programme functionality:

1. **Staffing**: a key, unavoidable, logistical challenge to programme delivery and high teacher engagement was staff absences due to staff sickness or school closures due to teacher strikes during the 2022/2023 school year. Teachers noted that absences caused momentum to be lost, and ultimately classes falling behind, including some schools not being able to complete the full ten weeks of the programme. Even in situations where cover teachers were in place, these staff were not necessarily familiar with, or able to access and deliver, the programme.
2. **Timetabling**: time constraints and fitting sessions into an already 'jam-packed' curriculum was a barrier to engagement. Teachers who discussed this in interviews and post-intervention focus groups explained that time pressures made finding three slots per week challenging, remarking that delivery of sessions often took longer than twelve minutes. In cases where teachers were critical of the amount of time programme delivery took, they described Stop and Think as 'becoming a chore' and delivery taking a significant 'chunk' of lesson time.
3. **Content**: teachers cited content inaccuracies or mismatch with current teaching practices. Those who gave examples of inaccuracies felt that they could introduce confusion to pupils: that pupils would struggle to understand the reasoning characters gave, with the teacher then needing to explain the intended misconception the programme was trying to convey. Teachers expressed concern that this could introduce 'new' misconceptions among pupils. One example was given for science: the visual content was of a person in space but the question related to a person stood on

the Earth. Content inaccuracies were, however, mainly mentioned in relation to maths. Examples for maths included:

- the programme referring to 'units' while current practice is to say 'ones';
- methods of multiplying and dividing by powers of ten not matching how this was taught in the classroom; for example, moving the decimal point versus moving the digits;
- maths answers not having the correct number of digits; for example, answer should have been '0.10' but was '0.1'; and
- fractions not being converted to their most simplified form resulting in pupils selecting the wrong answer.

4. **Technical problems**: teachers said technical issues with accessing the software resulted in increased delivery time, which then left less time for the rest of the lesson. A specific example was teachers having to make multiple attempts at loading the programme. Another was the game stopping halfway through and having to be restarted from the beginning rather than teachers being able to resume from where they had got to. Feedback from the delivery team suggests that these issues were related to connectivity with local broadband rather than something inherent in the game software.
5. **Programme functionality**: teachers discussed two specific issues. The first was not being able to conduct more than one session within the school day because of the timeout lock forcing a two-hour interval between sessions. This prevented teachers from easily catching up on missed sessions. The second was not being able to pause the programme. Teachers experienced this as frustrating as it prevented them from being responsive to pupil questions in real time and limited their ability to provide on-the-spot explanation and context to questions and answers.

Pupil engagement

Overall, we found variation in the levels of pupil engagement both between individual pupils and between schools. In the endline teacher survey, the majority of teachers (96%, n = 66) described pupils as either 'highly engaged' (57%, n = 39) or 'moderately engaged' (39%, n = 27), with 4% of teachers describing pupils as 'not very engaged'. In the qualitative research, teachers' accounts for different levels of pupil engagement related to pupils' different needs, in particular, pupils with SEND and pupils with varying academic ability.

High versus low engagement

In interviews and focus groups, teachers described what 'high' and 'low' pupil engagement looked like. Teachers reported that highly engaged pupils were excited when they saw Stop and Think on the screen at the start of lessons and disappointed when it ended. They described pupils identifying with the characters, cheering them on, and adopting the programme phraseology reminding each other to 'stop and think'. Across the lessons that we observed, we saw pupils focused on questions, answering questions when called upon, and offering explanations for answers. Pupils encouraged one another and applauded peers when their answer was correct. We heard directly from pupils in focus groups about their experiences playing. Overall, pupil accounts in focus groups were positive, with one pupil commenting it made them feel 'joyful and happy' to play the game.

Pupils displaying lower engagement levels complained about having to play the game and were more prone to being distracted or 'bored'—the latter due to frustration with the programme structure and losing interest due to its perceived repetitiveness. (The perceived ease or difficulty of the maths and science content is covered on page 79 in the subsection Maths Versus Science.)

Pupil engagement over time

In the endline teacher survey, we asked teachers whether pupils engagement changed or was consistent over time. The largest proportion of teachers (43%, n = 30) thought that pupil engagement 'stayed the same'; roughly equal numbers thought that pupil engagement had either increased (23%, n = 16), or the reverse, decreased (25%, n = 17). The remaining teachers (9%) thought pupil engagement varied over the programme duration.

Table 36: Whether pupil's engagement with the programme changed or stayed the same (endline teacher survey)

	Frequency	Percentage
Stayed the same	30	43%
Increased	16	23%
Decreased	17	25%
Varied	6	9%
Total respondents	69	100%

These results are interesting to consider in relation to the 96% (n = 66) of teachers reporting that pupils were engaged or very engaged and suggests that while some pupils' engagement levels varied over time, overall pupil engagement levels remained good.

Teachers commented about pupil engagement in more depth in interviews, focus groups, and the endline teacher survey. In cases where teachers described pupil engagement as consistent or increasing over the programme duration, they said this was because of (1) the novelty of each session with new topics and questions and (2) pupil confidence increasing over time as they became accustomed to the format and more confident in attempting questions.

In contrast, where engagement was felt to decrease this was put down to (1) the perceived repetitiveness of the programme, with pupils becoming uninterested towards the end of the programme and (2) technology issues such as lag time and frustration with glitches and not being able to correctly enter answers into the software. For example, a teacher focus group participant explained how initially their Year 5 pupils had been highly engaged but further into the programme they would complain about having to do it.

> *'I would say it's, "Right, we've got the Stop and Think today", and they'd go [groans], I'd be like, "Yes, and it's 12 minutes, we're just 12 minutes and it's done, but we need, we have to do Stop and Think." I think that they became less engaged as we went on' (teacher, post-intervention teacher focus group 2).*

Finally, teachers discussed differences between individual pupils' engagement levels within classes. In particular, they perceived that pupils with SEND were more likely to remain engaged for longer than peers without SEND.

Facilitators to pupil engagement

A variety of factors were named by teachers in both the endline teacher survey and qualitative research as contributing to high pupil engagement. First, teachers in interviews and focus groups reported that the programme generally aligned with pupils' academic ability. Teachers described this in terms such as Stop and Think being 'pitched in the middle'. These teachers felt that the programme provided enough challenge for more academically able pupils to engage while also allowing less academically able pupils, including those with SEND, to get at least some questions correct.

A second factor was the programme's graphics and design. Teachers noted that pupils identified with the characters' physical attributes, such as shared characteristics including ethnicity and hair colour. In focus groups, pupils complimented various programme visuals, such as the characters having a smile on their faces, and the backgrounds looking good:

> *'It has a positive tone and some positive stuff, like it has a bright sky, and that would probably motivate some people' (pupil, pupil focus group 1).*

Third, teachers who used the optional programme elements—the coins and leader board—described these as enhancing pupil engagement. Examples provided by teachers, in both the qualitative research and the post-intervention teacher survey, included pupils being invested in moving up the leader board and being keen to earn more coins to get a 'streak of coins'. In a class observation, we saw pupils using coins to change the avatar's outfit at the end of the game. A pupil from this class later commented, in a focus group, that their class had had discussions in the preceding week about how they could use their coins to change the avatar's outfit. In another group, pupils referred to the leader board and how while they enjoyed comparing their ranking with other schools, they would also like a 'local leader board' just for their school, to have a competition with other classes in their year group.

Teachers who had chosen not to use the optional programme elements explained they had decided against doing so as they did not want to introduce a competitive angle; in their view, this went against the ethos of the programme. These findings demonstrate that the elements being optional, rather than inbuilt and compulsory, works well and allows teacher choice over what will work best in their school for their pupils.

Fourth, teachers mentioned pupils' familiarity with using technology as facilitating engagement. They explained that since the COVID-19 pandemic, pupils have been using technology in education for several years, and it being 'just part of the equipment' in everyday usage.

Lastly, the game was considered the right length, with one teacher commenting:

> *'It isn't too long where it gets a little bit boring, and it isn't too short where you think, "I wanted more." I think it's just at the right length. I think if it was any longer, I think the children may go off it a little bit' (school lead 7, post intervention school lead interview).*

Barriers to pupil engagement

Three main barriers to pupil engagement were identified in the qualitative research with teachers. All of these distracted pupils' attention away from intellectually engaging with the content of the programme and broke their attention.

First, pupils and teachers discussed how they found malfunctions in the programme irritating and distracting. A variety were noted including seemingly incorrect question-and-answer routing, and problems with the smoothness of programme flow. Teachers described how pupils would track how often the programme was glitching, and this became a focus of attention rather than the game. The teachers that discussed what lower levels of pupil engagement looked like in the classroom suggested that 'glitches' (such as the wrong audio for question/answers, freezing on loading etc.) led to pupils becoming disengaged.

> *'I think that's maybe a glitch because we put the correct answer, but then [it didn't—then] it said we got it wrong. Then eventually, it showed us the correct answer, but that was the answer that we put!' (pupil, pupil focus group 1).*

Second, in pupil focus groups, a view was expressed that the structure of the programme was 'strange' and did not seem logical. One example that pupils gave was that after the first question, the programme routed to a bonus round even though only a single question (rather than a full round) had taken place. During a lesson observation, we observed a similar example of pupils critiquing the game programming: pupils noticed a pattern in the frequency with which avatars gave the correct explanation, which then made it easier for pupils to select their answers based on the observed pattern rather than intellectually engaging with the questions.

Finally, during a class observation we observed pupils pointing out spelling errors in the programme. After the observation, the teacher stated the game had spelling and grammatical errors that needed correcting, commenting this was obstructive to pupils' engagement and learning.

Repetition

Across interviews and focus groups, teachers and pupils reflected on repetition in the format of the questions in the programme. One view among teachers was that repetition was not a clear-cut barrier to engagement but rather discussed as being both a potential facilitator and barrier to pupil engagement—and to their own engagement in delivering the programme.

> *'I never got the sense that they were getting bored or restless with that … it's repetitive some of it … but also I felt they were engaging with it' (teacher, post intervention focus group 2).*

Teachers who discussed repetition as facilitating pupil engagement stated that the clear and standardised format of the programme could be reassuring for pupils. For example, pupils would excitedly try to beat the characters to saying catchphrases ('stop and think!'), and they enjoyed this element.

A contrasting view among teachers was that repetition was a barrier to pupil engagement. Teachers with this view described the programme as 'irritating', 'distracting', or 'boring' to pupils, which led to pupils 'switching off' from the programme. A particular example given by a teacher was the programme music having become monotonous to their class. In other instances, teachers discussed how although pupils had initially found the programme engaging, over time the repetitiveness of the programme eroded pupil engagement.

> *'Over time, their enthusiasm did dip as it became quite repetitive' (Endline Teacher Survey open text response to the question: 'Overall, how would you rate your experience of using the Stop and Think programme?).*

In cases where pupils discussed repetition in the programme, they did so in more straightforward negative terms that this made the programme 'boring'.

> *'It also keeps on repeating the same instructions, like, "Remember to stop and think, remember to stop and think." The same questions come up … it keeps on just repeating' (pupil, pupil focus group 1).*

Maths versus science

A view commonly expressed by both teachers and pupils was that pupils tended to find maths questions easier than science questions. Teachers noted that pupils with higher academic ability, or those in higher ability maths sets,[52] found maths questions 'too easy' and as a result did not engage as well with the game as lower academic ability peers. Pupil views chimed with this: in focus groups, they mentioned that many of the maths topics had already been covered in earlier school years, resulting in the questions being too easy.

Compared to maths, teachers felt that science questions were pitched at the 'right level' for all pupils. In addition, teachers viewed addressing science misconceptions as being particularly beneficial to enhancing pupils' subject knowledge. This finding may tie in with school leads telling us that a reason for signing up to the programme was the science component and the relative lack of existing science interventions compared to maths.

Pupil views about science questions paralleled teacher views: science questions were seen as more challenging than maths questions, as 'more fun', with more surprising answers. Pupils reported that, overall, they learnt more from the science questions than the maths questions. (Here, it is once again important to note that the intervention aims to increase pupils' inhibitory control skills rather than their subject knowledge in science and maths, though it is understandable that pupils would have approached the intervention as a knowledge gaining exercise.)

Year 3 versus Year 5

Across the qualitative data, we found some evidence that engagement levels varied by year group. Teachers specifically discussed that Year 5 pupils' engagement levels decreased over the programme duration while for Year 3 pupils engagement was more likely to remain consistently high. This chimes with the earlier finding that some of the content is currently too easy for pupils, particularly in the maths game.

Pupils with SEND

In the baseline survey,[53] teachers said that accessibility and inclusivity for all pupils could be a potential challenge to programme delivery. The qualitative research with teachers found a range of reflections about the accessibility and inclusivity of pupils with SEND.

One view was that Stop and Think was accessible and inclusive to pupils with SEND. Teachers who held this view described the programme as 'SEND friendly' and as providing opportunities for pupils with SEND to be academically challenged on an equal footing with their peers and build confidence in their own academic ability.

> *'The Stop and Think project gives them [pupils with SEND] that challenge within the questions sometimes, and that multiple choice, which sometimes they don't have access to … There's no differentiation when they're seeing that, they're all answering the same question. It's that where they see each other as an equal' (school lead 7, post-intervention school lead interview).*

Other examples of positive engagement among pupils with SEND included pupils who might normally be less inclined to participate in group activities joining in and 'having a go' at answering programme questions. Teachers reflected that the programme was well suited to the learning needs of pupils with SEND. This was because additional processing time

[52] In some schools in the qualitative sample, maths classes were organised into maths sets; the Stop and Think programme was delivered within this sets structure.

[53] We asked teachers about any challenges that could arise from participating in the programme: accessibility for all pupils was one factor identified alongside the (dominant concern) or time, and other factors—tech issues, pupil engagement, and using the programme effectively.

was an intrinsic component of the programme and it was not compulsory for pupils to verbalise answers if they did not want to.

> *'Everybody could take part. So, our children who struggle more in literacy-based subjects did really, really well with this. So, I've got two children in my class who are on the special needs register who loved it because they could join in … It was very inclusive' (school lead 1, post intervention school lead interview).*

A further interesting reflection was that the short, simple, and to-the-point questions were particularly accessible for pupils with SEND. In contrast, teachers thought that programme questions where additional steps were required were less engaging for pupils with SEND.

Included in teacher accounts of the accessibility and inclusivity of the programme was a caveat that pupils with SEND were able to engage with the programme with their peers so long as their usual SEND support arrangements were in place. Teachers also felt that achieving participation and retaining engagement for pupils with SEND could be challenging because of a range of reasons, such as a lack of confidence to participate in new activities, the need for additional processing time, and the need for clarification and guidance on what the question was asking. Specifically, teachers explained that they would reassure SEND pupils and encourage them to participate, ensure pupils slowed down, and provide additional explanation or reword questions.

Teachers also noted that, over time, engagement among SEND pupils could be prone to 'drop off' if they did not receive appropriate support. This was particularly noted in relation to question comprehension. Again, teachers highlighting this issue made specific mention of the need for support that met pupils needs as being essential to ensuring accessibility, inclusivity, and retaining pupil engagement.

> *'I found that there were a lot of times where I did just have to kind of explain or reword the question, where children with SEN were just struggling to understand it. … I think it definitely had to have additional input from the teacher to make sure that those children were accessing it as well' (teacher, post intervention focus group 1).*

Finally, while this topic was not widely raised, we found some evidence in the interviews with school leads and teachers that pupils with SEND engaged differently with maths and science questions. Specifically, they were initially more engaged with the science questions compared to maths. However, by the end of the ten-week programme, these pupils also built their confidence in relation to the maths questions.

Pupils with English as an Additional Language (EAL)

In interviews and focus groups with teachers, we asked how well pupils with EAL were able to engage with the programme. Overall, there was agreement that pupils with EAL were able to engage, although with the same caveat as for those with SEND: that they needed additional support to actively participate. A particular aspect of the programme highlighted was its highly visual format, especially in maths. This was felt to be important as large volumes of written text can be a barrier for pupils with EAL.

> *'We've received quite a few children who don't really speak English, or very, very little, so as it was quite visual, they could join in with that, which they did enjoy' (teacher, post intervention focus group 3).*

Differentiation

IPE RQ7 How different is Stop and Think to usual Key Stage 2 maths and science teaching?

This section discusses the teaching of Key Stage 2 maths and science across both control and treatment classes prior to and during the delivery of the Stop and Think programme. It explores the extent to which existing teaching methods for maths and science differ from Stop and Think. Understanding the difference between usual practice and the programme allows us to interpret the contribution that it has made in the observed outcomes from the impact evaluation and the perceived outcomes from the IPE.

Additionally, this section comments on the IPE dimension of monitoring the control and considers whether there was any risk of contamination in the trial. Since every school had one year group in the treatment group and the other year group in the control group, we are able to comment on monitoring the control by gauging any views around embedding Stop and Think into school-wide usual practices.

Usual approaches to Key Stage 2 maths and science teaching

Endline interviews with school leads and interviews with teachers highlighted that schools tend to follow a standard approach to teaching both maths and science at Key Stage 2. The majority of teachers described the following approach or something very similar:

- first, they introduce a new topic to the class and present examples in support;
- then, teachers link the topic to any other topics that they had covered in previous lessons;
- lastly, they provide worksheets to pupils and facilitate discussion in groups or in pairs with the aim of clarifying any misunderstandings.

In focus groups, teachers mentioned a variation of this approach that was used specifically for science lessons. Here, schools come up with an 'enquiry question' every half-term, for example, one relating to the life-cycle of a plant or the food chain. Teachers then complete retrieval tasks in the science lessons that link to the enquiry question, or they link the topics that pupils study in other classes to the science enquiry question.

In interviews, school leads described some additional approaches that are used in science and maths lessons. One was the 'Think, Pair, Share' strategy where pupils are given a chance to think of an answer first on their own, then share their answer with the pupil next to them and finally share it with the rest of the class. Another approach was to use Kagan structures[54] in teaching whereby pupils are given 'table points' for every correct answer that they give for questions, adding an element of competition within the classroom.

School leads were also asked about the use of technology in science and maths lessons. The use of Google Chromebooks, laptops, tablets, and interactive whiteboards was common in lessons and activities. These were sometimes supplemented with concrete equipment like models or physical resources so that pupils could understand concepts more comprehensively.

School leads and teachers also discussed differences between maths and science. For maths, several teachers mentioned in focus groups that they used the White Rose scheme,[55] which includes visual, hands-on apparatuses and representations as well as textbook exercises. White Rose was used to cover various topics like subtraction, multiplication, money, and time. Sometimes teachers supplemented the White Rose activities with additional challenges where pupils work independently on a few additional questions. One school lead interviewee mentioned using programmes like *Number-Sense*, which included 15-minute animation videos about different mathematical concepts, and *Time-Tables Rock Stars*, where pupils played games individually or against each other to learn maths.

Compared to maths, school leads described science as a subject where pupils learn through investigation and experiments rather than just from textbooks. One view among teachers in focus groups was that science lessons were, at times, connected to what they were doing in English or other lessons so that pupils could form links between different subjects. Additionally, teachers use science scheme guides to help them decide what topics to cover at which points in the term. Lastly, some schools conduct dedicated science days every half-term where they review topics that they have covered during the term using activities like exploring nature and making science models and so forth.

Teachers highlighted some variations in the time that they spent preparing for maths compared to science lessons. We asked teachers in the pre-intervention survey about time used for lesson preparation: for maths, they reported an equal split between spending 16 to 20 minutes, 21 to 30 minutes, and 31 to 59 minutes. For science, most teachers (34%, n = 53) reported spending 21 to 30 minutes followed by 31% (n = 48) stating that they spent between 31-59 minutes. Interestingly, 11% (n = 17) of teachers stated that they had spent between one and two hours preparing for science lessons, which was a much higher figure compared to typical maths lessons.

Use of computer programmes

The use of computer programmes varied substantially between different schools and for different subjects. On the whole, the pre-intervention survey found that computer programmes were more widely used in maths lessons compared to science: 42% (n = 66) of teachers mentioned that they had used computer programmes for *maths* compared to 11% (n = 17) for *science*; 26% (n = 41) reported not having used any computer games in their teaching Interestingly, when asked whether they had used computer programmes for subjects like reading or English in the past academic year, a

[54] Kagan S, (2011). The "P" and "I" of PIES: Powerful principles for success, Kagan Publishing, Kagan Online Magazine.

[55] https://whiteroseeducation.com/resources

majority of teachers (67%, n = 105) indicated that they had done so, meaning that science computer programmes were a relative outlier in this regard. The use of computer programmes as a teaching tool did not substantially differ between the control group and treatment group teachers.

Of the 26% that reported not using any computer programmes, teachers presented a wide range of reasons for this: 53% (n = 27) reported that they had not considered using computer programmes to teach maths and science. This was followed by 20% (n = 10) reporting that there was a lack of suitable equipment to set up computer programmes in their school. Other reasons stated by teachers included programmes being prohibitively costly (18%, n = 9), while a handful (5%, n = 5) felt that they were not confident in using computer programmes in their teaching.

The use of computer games as a teaching tool did not substantially differ for the control group teachers. The pre-intervention survey found that almost half of the control group teachers (47%, n = 23) had used computer programmes for Key Stage 2 maths while 22% (n = 11) had used computer programmes for both science and maths. They particularly mentioned using computer programmes like *TT Rockstars* and *Hit the Button*. However, 25% (n = 12) of control group teachers reported not having used any computer programmes in their maths and science teaching. More than half of the control group teachers (69%, n = 33) reported that they had used computer programmes in lessons outside of maths and science; this is similar to the finding from the treatment group teachers.

In interviews, teachers mentioned using computer programmes in their maths lessons quite frequently. They highlighted maths programmes that focused on multiplication and fractions, for example, *Hit the Button* and *Multiplication Check*. A vast majority stated that they did not use any computer programmes for science and focused on other 'investigative' and practical activities instead, like going outdoors, making torches, and the like. There was also a view among school leads that maths game apps and multiple-choice game software were much more common when compared to games for science.

The views of pupils corresponded with those of teachers and school leads. In focus groups, one group of pupils said that they had never used computer programmes in their usual practice maths and science lessons. Another group mentioned using computer programmes for maths specifically, such as *TT Rock Stars*, *Hit the Button*, and a platform called *Mangahigh* that included specific games for integers and fractions. There was consensus among pupils that they had made very little use of any computer programmes for science. A handful mentioned using software where they could identify animals for their science classes, but they had not used it many times during lessons. Overall, a strong view among pupils was that using computer programmes made lessons more enjoyable. Even pupils who had not used computer programmes were keen to use laptops or iPads within their maths and science lessons.

Counterintuitive concepts in usual practice

In the pre-intervention survey, treatment and control group teachers were asked whether they use any particular strategies for teaching counterintuitive concepts in maths or science. Based on these responses, counterintuitive strategies were more commonly used for teaching maths than science: 36% (n = 36) mentioned that they had used these strategies for maths lessons, while only 20% (n = 31) reported using them for science. Although the difference in these proportions is relatively small, they provide a modest indication that counterintuitive strategies may be less commonly used in science compared to maths. It is also worth noting that teachers' awareness of counterintuitive concepts may have been influenced by their participation in the Stop and Think trial. This limits our understanding of teachers' knowledge prior to being recruited to taking part in the evaluation.

In order to teach counterintuitive concepts, treatment group teachers discussed using strategies like making deliberate errors so that pupils could correct them during maths and science lessons. In interviews, school leads affirmed that making mistakes was encouraged as a part of their usual teaching as this helped pupils gain a better understanding of where they might be going wrong. Additionally, teachers used visual resources and retrieval questions to help with learning. Teachers in the pre-intervention survey listed using games and programmes, visual modelling, revisiting topics, and other media to help explain these concepts in more detail.

For science, school leads from the treatment group expressed that they had used programmes like *Concept Cartoons*, which asks pupils to think before answering questions in a similar way to Stop and Think, as well as *Explorify*, which asks pupils to name what they see on the screen (like plants, animals etc.). They said that these programmes for counterintuitive concepts had highlighted that many teachers were already quite familiar with misconceptions in their usual lessons and were actively trying to support pupils with learning about misconceptions.

How Stop and Think differs from usual practice

In the IPE, we asked teachers and pupils to compare Stop and Think to their usual teaching practice. Here, we want to remind the reader that the intervention aims to improve inhibitory control skills in children, not to increase their subject knowledge in science and maths. Teachers' and pupils' responses below show that it may have been difficult for them to view the programme through the lens of improving inhibitory control skills rather than how lessons are structured and taught; perhaps understandably, considering the novelty of a programme targeting inhibitory control skills.

Teachers' perceptions of differences between usual lessons and the game

In the post-intervention survey, we asked treatment teachers whether Stop and Think is different to their usual practice when teaching science and maths: only about a third (31%, n = 22) reported that it was different from usual maths lessons; for science, around half (55%, n = 38) reported that it was different. For both subjects, the teachers who thought that the game was *different to usual teaching* explained that questions were presented differently to how they would be posed within a usual lesson and that the game was based on interaction and did not focus on actual problem-solving skills. Those who thought the game was *similar to usual teaching* explained that this was because the game used similar visuals as their textbooks, that the content of the curriculum was the same, and because pupils were generally familiar with the content.

In interviews with school leads and teachers, we also found different views about how similar or different the programme is to usual practice. A common view was that their usual lessons were similar: teachers explained that both the game and their lessons were set up in a way where a topic is introduced, more explanation is given about it, and pupils are asked questions. They referred to this way of conducting a maths lesson as the National Centre for Excellence in Teaching Mathematics (NCETM) approach.[56] In their view, the main difference was that usual practice lessons are more focused on the completion of worksheets, sometimes in a collaborative way among pupils (either in pairs or small groups) or individually with help from the teachers. By contrast, the Stop and Think game asks pupils to think about their answers individually and then discuss them with the teacher as a whole-class activity. Similarly, school leads observed that the concepts from Stop and Think were not new to them, however, the game acted as an extension to existing ways of teaching.

One view among teachers was that the game did not include many visuals compared to their usual science and maths lessons that used schemes such as *White Rose* wherein pictorial and concrete prompts and figures are used to help pupils understand challenging and abstract maths concepts. These teachers suggested that visuals should be more prominent in the game and should be added into every question, not just the science ones. They were of the view that the game sometimes felt quite 'arithmetic' and visual aids could help fix this issue. School leads expressed a different view—that the visual element of the Stop and Think game represented an innovative way to address misconceptions in pupils. For them, Stop and Think was a good addition to lessons as it was not always feasible for teachers to set up a visual component. (Here, it is worth noting that not all school leads were involved in classroom teaching.)

Pupils' perceptions of differences between usual lessons and the game

In focus groups, pupils reported that, on the whole, the Stop and Think game was different to their usual way of learning maths and science. A key difference was that the game was shorter than their usual lessons, which could last up to 30 minutes or more compared to the game's ten minutes.

For maths, one view among pupils was that the game did not require them to work out the answers and they could give answers directly: in usual maths lessons they were required to show how they worked out their answers in their worksheets or on the whiteboard. Another group of pupils said that they enjoyed Stop and Think maths lessons more because it followed the format of a game. One view among pupils was that usual maths lessons were more difficult than the Stop and Think maths questions; this was mainly because pupils reported that they had already studied the maths topics covered in the game which made it easier to play.

For science, pupils mentioned that usual lessons involved many different types of activities like measuring shadows, planting beans, making torches, labelling and drawing skeletons, and so on. A strong view among pupils in focus groups was that usual science lessons were more interactive and more task-based compared to maths lessons. Overall, pupils mentioned that they enjoyed their usual science lessons equally as much as the Stop and Think science game.

[56] https://ncetm.org.uk/about-the-ncetm/

Moreover, pupils discussed that topics within the Stop and Think game could vary in difficulty with some questions and topics being easier than the others. They highlighted that it was easier for them to answer Stop and Think questions that had already been covered in their usual maths or science lessons. This meant that some pupils found the game less challenging when compared to their usual maths or science lessons, especially for the maths questions. This view was also shared by school leads who had observed that the maths questions were less challenging in the game.

Whether the game led to changes in usual practice

In the pre-intervention survey, 26% (n = 41) of teachers expressed a desire to use computer programmes for maths and science, while 30% (n = 47) wished to use them for maths only. However, a large proportion of teachers (43%, n = 67) reported that, for the current academic year, they did not have any plans to use any computer games to teach maths and science. This difference in actual usage could link back to the reasons around prohibitive costs and lack of knowledge of which games to use discussed by teachers both in focus groups and in the surveys.

In interviews, intervention group teachers were asked to reflect on whether they had made any changes to their maths and science teaching as a result of Stop and Think. A majority of teachers reported that they had not done so. One reason for this was that they were already aware that giving more time to pupils to think through their answers was beneficial: as school leads noted, this practice was already being implemented using the 'five-second wait time rule' (not as long as the minimum ten-second wait in the Stop and Think game). According to teachers, other key elements of the game were also already part of teachers' existing practice, such as addressing misconceptions and the focus on repetition, which mirrored an existing focus on recall and retrieval within their school.

Here, we want to caution the reader that the teachers' and school leads' perception of what they knew before taking part in Stop and Think may have been influenced by their ongoing involvement in programme delivery at the time of the interviews. We did not conduct any pre-intervention interviews with teachers and school leads, and even if we had, participants would have already been recruited for the evaluation by that point and therefore had awareness of the need for pupils to 'stop and think'. (We did ask questions about teachers' awareness of counterintuitive strategies in the pre-intervention survey but these findings are similarly limited by the fact that teachers had already been recruited for the evaluation and in many cases had already received the Stop and Think training. The survey findings are discussed earlier in this chapter.) Because of the challenge in measuring teachers' genuine pre-existing knowledge of inhibitory control, we urge caution in interpreting these findings.

Another view among teachers and school leads was that the game had helped them remind pupils to stop and think, particularly in instances where they could see a pupil rushing through their answers. In one example, a teacher noted that they had started to ask pupils to 'stop and think' outside the Stop and Think sessions, and pupils would count ten or fifteen seconds before giving their answers to any questions. Interestingly, this was reflected in the pupil focus groups where pupils mentioned that the game had made them realise that being the fastest was not important. Pupils with this view reflected that answering correctly was more important than writing down an answer without thinking.

Teachers also expressed an intention to continue using both the game and their usual lessons beyond the trial, if possible: they suggested using the game in the future for usual maths or science lessons as well as with different classes. They mentioned that it would be beneficial to use the game alongside their usual paper-based activities and other computer games to ensure that there was a variation in the ways in which pupils learned maths and science.

One finding from the teacher focus groups was that even if a teacher was interested in using Stop and Think in their usual lessons in the future, they would not be able to do so without approval from the Senior Leadership Team (SLT). Teachers mentioned that the SLT was focused on wider strategies like 'Think, Pair, Share' or the *White Rose* method and did not want to give those up in favour of newer programmes like Stop and Think.

All teachers in interviews said that they had continued using their usual approaches to teaching maths and science alongside the Stop and Think game, even those who had made some changes to their teaching following the programme. They also said that they did not plan on making any substantial changes to their usual approaches. However, one view discussed by teachers in focus groups was the intention to include more interactive elements into their teaching in the future.

Finally, school leads reported that the Stop and Think game fits well with their schools' learning objectives. These school leads expressed a desire to continue using the game in the future. However, they found it difficult to comment on whether teachers would incorporate the game into usual lessons in the future. This was because teachers had only played it for a short period of time during the trial.

Monitoring the control

Every school had one year group in the treatment group and the other year group in the control group. In interviews with school leads and focus groups with teachers, we asked them to comment on whether there had been any sharing of information between treatment and control group classes at their schools to assess contamination. Teachers and school leads either did not comment on this or were unsure if learning from Stop and Think was shared among classes. In general, they were more certain that the game was not played by classes that were not meant to play the game. However, they did not comment on whether teachers in their school had discussed the game with other classes or used learnings from the game in their usual lessons with other classes. In interviews, teachers gave accounts of mixed year group classes playing the games together, but these were limited to either Year 3 and Year 4—or Year 5 and Year 6—playing it together. There are no accounts of the game being played by both Year 3 and Year 5 at the same school.

Perceived outcomes

IPE RQ8 What outcomes do teachers and pupils perceive to result from Stop and Think?

This section reports on perceived outcomes of the programme for teachers and pupils. The outcome areas we discuss are pupils' ability to learn counterintuitive concepts and common misconceptions, and one short-term intended outcome of reduced impulsive responding on maths and science questions from the Stop and Think logic model (Figure 2, page 11). The section concludes with a discussion of unintended outcomes.

Short-term outcome—reduced impulsive responding on maths and science questions

At baseline, almost eight in ten teachers (79%, n = 123) reported pupils generally 'rushed through' or responded impulsively when answering maths and science problems. Nearly all (98%, n = 153) thought this caused pupils to make more mistakes.

At endline, teachers were asked whether as a result of the programme pupils had reduced their impulsive responding on maths and science questions. Around one in four (25%, n = 17) thought it had been impactful in this regard for maths questions; a higher percentage (28%, n = 19) felt it was the case for science questions.

Table 37: Whether teachers agree with the statement: 'Since playing Stop and Think, pupils have reduced impulsive responding on maths questions' (endline teacher survey)

	Frequency	Percentage
Agree	17	25%
Disagree	14	20%
Neither agree nor disagree	28	41%
Unsure	10	14%
Total respondents	69	100%

Table 38: Whether teachers agree with the statement: 'Since playing Stop and Think, pupils have reduced impulsive responding on science questions' (endline teacher survey)

	Frequency	Percentage
Agree	19	28%
Disagree	13	19%
Neither agree nor disagree	27	40%
Unsure	8	12%
Total respondents	67	100%

In open-text responses, those who agreed that pupils' impulsive responding had reduced explained that pupils now paused more before answering and re-read questions to consider their meaning. For science specifically, teachers mentioned pupils using more scientific vocabulary in their answers. By contrast, teachers who disagreed with the statement stated they had not seen a change in pupil behaviour, with the majority of pupils still 'jumping feet first into answers'.

In interviews and focus groups, teachers reasoned that it was difficult to assign these impacts directly to the programme. In one example, a teacher explained they had already been instilling 'stop and think' behaviour in pupils before the intervention. Others explained that while having Stop and Think as a standalone activity may have helped to reinforce the message of thinking before responding, other factors, such as the academic ability of the pupils and the specific question being asked, made it difficult to determine the programme's impact at the class level:

> *'I tried to give them a bit more time to actually stop and think about the question even before Stop and Think was coming to us … It also depends on which children we have in front of us' (school lead 1, post-intervention school lead interview).*

In focus groups, pupils expressed mixed views about whether their impulsive responding had reduced. One view was that the programme had been effective in this regard. These pupils cited how taking more time to respond to maths questions had improved their chances of getting the correct answer:

> *'I'm usually quite good at maths but I sometimes get questions wrong because I don't read the actual question right. I skip ahead and miss part of a question and Stop and Think has really helped me improve that' (pupil, pupil focus group 1).*

A group of pupils also reported that Stop and Think had changed their assumptions about the speed with which schoolwork needed to be completed. This had challenged the notion that faster completion corresponded with higher intelligence:

> *'Before Stop and Think I just tried to do the fastest out of my class, but now I don't think being the fastest is the most important' (pupil, pupil focus group 2).*

A contrasting view was that the programme had made no difference to pupils' impulsive responding. One explanation pupils gave was that the effect was limited to when they were playing the game and did not translate to other contexts. Teachers agreed with this observation, with one explaining that six weeks post programme,[57] it was no longer at the forefront of pupils' minds and that they had stopped spontaneously using the phrase 'stop and think' in other lessons. Pupils and teachers also said that the maths games did not require any consideration time as the questions were not challenging enough.

Pupils' ability to learn counterintuitive concepts and common misconceptions

In the baseline teacher survey, almost all (99%, n = 155) teachers reported that pupils found maths and science problems containing counterintuitive concepts and common misconceptions challenging. In the post-intervention survey, we asked teachers whether pupils' ability to learn counterintuitive concepts in maths and science had improved since playing Stop and Think. These results are presented in Table 37 and Table 38. The survey asked teachers separately about maths and science. Here, most teachers (59%, n = 40) neither agreed nor disagreed or were unsure about whether the programme had improved pupils' ability to learn counterintuitive concepts in maths; the corresponding figure for science was lower at 53% (n = 37) of teachers were either unsure or they neither agreed or disagreed that the programme had improved pupils' ability to learn counterintuitive concepts in *science*.

[57] Teacher focus groups took place between May and July 2023.

Table 39: Whether teachers agree with the statement: 'Since playing Stop and Think, pupils have improved their ability to learn counterintuitive concepts in maths' (endline teacher survey)

	Frequency	Percentage
Agree	22	32%
Disagree	6	9%
Neither agree nor disagree	27	39%
Unsure	14	20%
Total respondents	69	100%

Table 40. Whether teachers agree with the statement: 'Since playing Stop and Think, pupils have improved their ability to learn counterintuitive concepts in science' (endline teacher survey)

	Frequency	Percentage
Agree	26	38%
Disagree	6	9%
Neither agree nor disagree	25	36%
Unsure	12	17%
Total respondents	69	100%

Teachers were given the option to add additional explanation in open-text responses. Those who thought their pupils' ability to learn counterintuitive concepts in *maths* had improved highlighted pupils' increased engagement in learning. Others said the programme simply complemented existing teaching approaches rather than directly effecting pupils' ability to learn counterintuitive concepts. Among the positive responses about *science*, teachers mentioned that the programme had enhanced pupils' receptiveness to learning science in general, including outside the Stop and Think sessions. Others thought that any improvement was simply due to pupils doing more science, rather than specifically playing the game.

In the qualitative research, teachers expressed a view that the programme mainly helped a subset of pupils with maths misconceptions—those working at or below expected academic levels—whereas the programme helped all pupils with science misconceptions:

> *'I think it has helped some of their maths misconceptions for those children, that group of children. It's helped the whole cohort for science' (intervention teacher, school visit 2).*

Another view among teachers was that the programme sometimes introduced 'new' misconceptions to pupils in maths that could lead to confusion. One example concerned content about multiplication and the programme detailing a method that contradicted how pupils had been taught.

> *'It didn't match up with the way that we would deliver that … the rules that we would follow in class. It actually ended up introducing some misconceptions to how we would teach it, which was obviously a bit disappointing (intervention teacher, post intervention teacher focus group 1).*

To mitigate against confusion and 'new' misconceptions developing among pupils, these teachers wanted the programme to allow them to pause a session to provide further information and clarification.

Improved attainment on maths and science questions

In the post-intervention survey, around half the teachers thought the programme had improved pupils' attainment in maths and science. There was a small difference between science (51%, n = 35) and maths (41%, n = 28).

Table 41. Whether teachers agree with the statement: 'Stop and Think has supported pupils to improve their numeracy skills' (endline teacher survey)

	Frequency	Percentage
Agree	28	41%
Disagree	13	19%
Neither agree nor disagree	21	30%
Unsure	7	10%
Total respondents	69	100%

Table 42. Whether teachers agree with the statement: 'Stop and Think has supported pupils to improve their science skills' (endline teacher survey)

	Frequency	Percentage
Agree	35	51%
Disagree	8	12%
Neither agree nor disagree	18	26%
Unsure	7	10%
Total respondents	68	100%

In the open-text responses, teachers who agreed with the statement for both subjects said Stop and Think had helped to reinforce concepts and cement learning among pupils and that the content was well-aligned with the curriculum. For *maths*, the programme had increased pupils' confidence to attempt questions; for *science*, it had exposed misconceptions. Those who disagreed said that the programme did not add anything to a usual good lesson and that the level of maths in the programme was more appropriate to lower-ability pupils.

We also asked teachers and pupils this question in the qualitative research encounters. Teachers did not feel able to judge whether the programme alone had improved pupil attainment. There was, however, a view among teachers that the programme was more useful in improving pupils' science rather than maths attainment. Pupils on the whole were more positive: they discussed how the programme had helped them improve in both science and maths and that they were now able to answer more challenging questions correctly.

Differential impacts on pupils

In the endline survey, a third of teachers (32%, n = 27) reported that *all* pupils benefitted equally from playing Stop and Think. However, a sizeable percentage (22%) said that pupils with SEND and, according to 15%, pupils with EAL particularly benefitted from playing the game; 18% (n = 15) of teachers were unsure about whether specific groups of pupils had benefitted. (Note that the sample sizes are small.)

Table 43: Whether teachers felt that certain groups of pupils benefitted from playing Stop and Think (multiple response question in the endline teacher survey)

	Frequency	Percentage
Lower socio-economic status pupils	5	6%
Pupils with special educational needs and disabilities (SEND)	19	22%
Pupils with English as an additional language (EAL)	13	15%
Other group of pupils	6	7%
All pupils benefitted equally	27	32%
Unsure	15	18%
Total (number of responses)	85	100%

Similarly, in the qualitative research, we found a range of views from teachers about whether different groups of pupils benefitted from the programme more than others. One view was that pupils with SEND particularly benefitted from Stop and Think: teachers described how pupils with SEND usually found it challenging to access lessons and to maintain focus but that they were able to do so during Stop and Think sessions. In one example, a teacher reflected on a pupil with ADHD in their classroom. At the start of the programme, the pupil would shout out answers but by mid-point they were able to focus on the game. This improved learning behaviour had extended outside the Stop and Think sessions.

Another view expressed by teachers was that the programme particularly benefitted pupils with lower academic ability. Teachers explained these pupils were more likely to find the programme content challenging whereas more academically able pupils more commonly found the game 'boring' as they knew the answers without needing to stop and think. These teachers highlighted that this was particularly the case with the maths content. In contrast, science content was perceived to benefit *all* pupils with misconceptions in science, no matter what their ability.

Teachers on the whole did not discuss the programme as being particularly beneficial for those from lower socio-economic backgrounds.

Unintended outcomes

Alongside the perceived benefits directly linked to the Stop and Think logic model, teachers and lead teachers also discussed several unintended outcomes.

Firstly, teachers in interviews said that taking part in Stop and Think had led to them reflecting on their own practice. This was particularly in relation to science. Teachers found the programme had improved their science knowledge and enabled them to more effectively assess where pupils needed help with misconceptions and to plan teaching accordingly. Examples included one teacher looking up information where they had not themselves known the correct answer to a science question, and another more regularly bringing up misconceptions in their science lessons.

Secondly, the programme supported a more inclusive classroom culture. Teachers described the programme design as being accessible to all because it allowed anyone to answer and it did not matter if they got the answer wrong. This was perceived to increase pupil confidence to attempt any question.

> *'[The programme] makes pupils braver to get involved: anybody can have a go' (intervention teacher, school visit 4).*

The programme also reduced the stigma associated with different pupils needing different amounts of processing time. One example was pupils praising their peers with SEND when they volunteered answers during the game.

Regarding the point about inclusivity, it is worth noting that in our interview with the BIT delivery team we did hear of a school deciding that a set of lower academic ability pupils would *not* take part in the programme. This was on the basis that programme was perceived to be 'too challenging'. This indicates that the potentially beneficial effects of the programme for inclusivity and classroom culture may not be fully appreciated by all schools.

Cost

Table 44 summarises the added costs associated with delivering Stop and Think incurred both by schools (compared to usual teaching practice) and by the delivery partner (BIT). For schools, the activities covered represent the added effort or time of implementing the programme over and above teachers' usual responsibilities. Following EEF guidance (EEF, 2023) and the 'ingredients method' (Levin, et al., 2018), we have presented the following cost categories, including personnel costs for:

- preparations for programme delivery (BIT);
- training for implementation of the programme (schools and BIT); and
- implementation of the programme (schools and BIT).

We have also included facilities, equipment, and materials required for implementation.

We acknowledge the EEF's guidance to produce the headline estimate of costs in line with the primary outcome analysis, which in our case would be figures for FSM pupils allocated to maths testing. However, NPD data including FSM status was not accessed for the whole sample of pupils enrolled in the study, hence we only present figures based on the overall sample.

Table 44: Cost table

Category	Cost ingredient/activity	Cost incurred to	Personnel group	Start-up or recurring cost	Time required (hours)	Estimated cost (5% VAT included)
Personnel costs for preparing programme delivery	Recruiting S&T assistants, developing training materials for teachers and assistants, and designing the system for scheduling school visits	BIT	BIT core staff	Start-up	418.58 h	£37,071.82
	Delivering training to S&T assistants and scheduling school visits	BIT	BIT core staff	Start-up	270.58 h	£20,438.30
	S&T assistants being trained and scheduling school visits	BIT	S&T assistants	Start-up	512.00 h	£9,539.20
Personnel costs during training for implementation of the programme	S&T assistants delivering training to school staff	BIT	S&T assistants	Start-up	1,784.25 h	£46,632.97
	Teacher cover during training	Schools	School staff	Start-up	73.75 h in total in the trial (0.25 h per teacher, 0.43 h per school)	£2,175.56 in total in the trial (£7.28 per teacher, £12.58 per school)
Personnel costs for the implementation of the programme	Teacher time preparing first ST lesson	Schools	School staff	Start-up	54.6 h in total in the trial (0.18 h per teacher, 0.32 h per school)	£1,610.74 in total in the trial (£5.39 per teacher, £9.31 per school)
	Teacher time preparing S&T sessions after the first session, additional time delivering lessons with S&T	Schools	School staff	Recurring	2,172.04 h in total in the trial (7.26 h per teacher, 12.56 h per school)	£64,076.83 in total in the trial (£214.30 per teacher, £370.39 per school)
	Managing the implementation of the programme	BIT	BIT core staff	Recurring	823.42 h	£66,214.09
	Helpline assistance and ongoing support to schools	BIT	S&T assistants	Recurring	886.00 h	£19,052.54
	Developer set-up costs, assistance and support (for example, in case of any questions about the software)	BIT	Birkbeck	Start-up	n/a	covered by 2x £30,000 payments (BIT to Birkbeck)
Facilities, equipment, and materials for implementation	Computer and a projector or an interactive whiteboard to deliver the sessions (requirement for recruitment to the intervention)	Schools	n/a	Pre-requisite	n/a	n/a
	Printing teacher handbooks, arranging DBS checks for four S&T assistants, obtaining HubSpot licenses	BIT	n/a	Start-up	n/a	£3,119.00

Teacher time cost

A total of 69 teachers responded to our post-intervention teacher survey and reported that they used Stop and Think. The following time cost estimates are based on their survey responses. The cost evaluation assumptions and analyses are based on teacher costs (does not include TA costs).

Teacher start-up costs

Attending training

Six teachers reported that they did not attend the training. Those who did (n = 63) spent on average 48.33 minutes in training, with teachers most commonly reporting having spent 46 to 60 minutes (n = 27) or 31 to 45 minutes (n = 15). To avoid costing twice for teacher time during training, we only include the teacher cover required during training (see below) in teacher cost calculations (EEF cost guidance, 2023).

Teacher cover during training

Most teachers reported that teacher cover was not necessary while they were in training (n = 48). Some reported it was required for 46 to 60 minutes (n = 8). Among those who said teacher cover was needed (n = 19), the average was for 52.18 minutes. Overall (n = 67), teacher cover was needed for an average of 14.8 minutes.

Time preparing first Stop and Think lesson

Teachers spent on average 10.96 minutes preparing for the first Stop and Think lesson (n = 69), including one who reported not spending any time doing this. This has been reported as zero minutes spent in our calculations. Most commonly, teachers reported spending one to five minutes (n = 28) or six to 15 minutes (n = 22) to prepare.

Recurring teacher costs

Time preparing after the first Stop and Think lesson

After the first Stop and Think lesson, teachers spent on average 3.97 minutes preparing the sessions (n = 68), including three who reported they did not spend time doing this. Most commonly, teachers spent one to five minutes preparing (n = 57).

Additional time delivering a lesson with Stop and Think

Fifty teachers reported that Stop and Think increased lesson delivery time, while 16 reported that it did not. The reported time Stop and Think added to lessons was on average 14.33 minutes (n = 47), with teachers most commonly reporting it added six to 15 minutes (n = 24) or 16 to 20 minutes (n = 22). Overall (n = 63), it added an average of 10.69 minutes to lesson delivery time.

Average salary

In order to translate expended teacher time into a monetary estimate, we collected teachers' annual salary in the teacher survey. As a reminder, we explore only the *added* cost of implementing the programme, over and above usual teacher responsibilities. Thirty-one teachers provided their annual salary, which averaged at £37,318. Estimates of monetary teacher cost provided below are based on this average annual salary. To estimate hourly pay rates from average reported annual salaries, we assume a maximum of 1,265 directed hours completed by teachers in a year. This produced an average hourly pay rate of £29.50, which we use to calculate monetary cost of teacher time below.

Estimating per teacher cost

The average additional time associated with delivering 30 sessions of Stop and Think per teacher is 25.76 minutes (£12.66) for start-up time costs (that is, teacher cover during training, time preparing first Stop and Think lesson) and 7.26 hours (£214.30) for recurring time costs (that is, time preparing Stop and Think lessons, additional time delivering a lesson with Stop and Think). This amounts to 7.69 hours (£226.96) per teacher.

Following EEF guidance, we further calculate projected per-teacher costs associated with delivery over three years. For the first year, both start-up and recurring teacher costs are included in the estimate; only recurring teacher costs are included in the estimate for the two subsequent years. Calculations assume 30 sessions are delivered per class in a year, and one teacher per class delivering Stop and Think.

The estimated cost of teacher time for delivering Stop and Think across 3 years amounts to 7.69 hours (£227) per teacher in the first year and 7.33 hours (£216) per teacher for each subsequent year. Over three years, this equates to 22.35 hours of teacher time and an estimated £659.48 pay per teacher, with an average of 7.45 hours and £219.83 pay per teacher per year over three years.

Estimating teacher cost per school

Estimates of teacher cost per school are based on 299 teachers delivering Stop and Think across 173 schools in the trial. In the trial, 299 teachers spent an estimated 2,300 hours delivering the programme across 173 schools (that is, an average of 13.30 hours of teacher time per school), including start-up and recurring costs. Using average salaries provided, the monetary cost of this across all schools in the trial is estimated at £67,863—an average of £392 per school.

Estimated costs of teacher time per school for delivering Stop and Think across three years involve on average a total of 39 hours (£1,140) of teacher time per school, at an average of 13 (£380) hours each year.

Delivery partner cost

We obtained start-up costs for preparing programme delivery from BIT as the delivery partner. The three cost ingredients presented represent three stages of preparations that BIT were able to estimate time resources for, rather than a split between specific activities (see Table 42). BIT also provided estimates of recurring personnel costs during the implementation of the programme, including implementation management, helpline assistance, and ongoing support to schools as well as start-up costs for materials and prerequisites for delivery.

BIT further provided costs associated with developer set-up, assistance, and support by Birkbeck during the implementation of the programme. For this cost ingredient, it is not possible to differentiate start-up from recurring cost activities. Hence, we assume this cost ingredient to be start-up rather than recurring for the calculations below as this includes the cost of setting up the programme. Additionally, developer involvement with delivery can be expected to decrease after initial implementation, especially with the support offered to schools by the delivery partner.

Estimating total programme cost

Following EEF guidance, we summarise the costs associated with delivering Stop and Think as if delivered across three years, with both start-up and recurring costs included in the estimation for the first year, and only recurring costs included for two subsequent years (EEF guidance). We include both costs incurred to schools and to the delivery partner in this calculation.

In order to estimate monetary cost of teacher time, we use average provided salaries in the endline teacher survey (see above) and the total number of teachers and schools in the trial. As mentioned above, we categorise delivery partner costs during preparation for the delivery of the programme, as well as developer costs for assistance and support during implementation, as start-up costs for the purpose of the following cost estimations.

The total cost associated with delivering Stop and Think in the first year (that is, including start-up and recurring costs to schools and the delivery partner) is £329,931, consisting of £180,587 total start-up costs and £149,343 recurring costs. The total for delivering each subsequent year (that is, only including recurring costs to schools and the delivery partner) is £149,343. The resulting total cost across all three years is £628,618, with a per year average of £209,539.

Cost per pupil

A total of 14,718 pupils received Stop and Think in the trial. Hence, the cost of delivering Stop and Think per pupil per year across 3 years is £42.71 in total over three years, with an average of £14.24 per year.

Conclusion

Table 45: Key conclusions

Key conclusions
1. Pupils eligible for FSM receiving Stop and Think made no additional months' progress in maths attainment compared to FSM pupils receiving teaching as usual. This result has a moderate to high security rating. All pupils receiving Stop and Think made no additional months' progress in maths attainment compared to pupils receiving teaching as usual.
2. All pupils receiving Stop and Think made two additional months' progress in science attainment compared to pupils receiving teaching as usual. FSM pupils receiving Stop and Think made one additional month's progress in science attainment, compared to FSM pupils receiving teaching as usual. Further analysis suggests that the impact on FSM-eligible pupils was similar to the impact for all pupils in science.
3. The Stop and Think programme largely took place as intended with most participating teachers delivering the 30 intervention sessions. However, due to scheduling and staffing issues, some teachers could not always follow the model of three sessions per week at the start of maths and/or science lessons within the ten-week period.
4. Evidence to support the short-term outcomes of reduced impulsive responding and improved attainment on science and maths misconceptions tests is inconclusive due to the low reliability of the assessment methods used. No link was found between socio-economic status (measured by FSM), increased participation, or compliance with Stop and Think and improved maths attainment.
5. Some teachers and pupils perceived the maths content to be easier in comparison to the science content. There is evidence to suggest the programme also showed only minor differences from usual maths and science teaching practice. This perception may have been because they saw the programme as an extension of their usual science and maths teaching rather than a programme for enhancing pupils' inhibitory control.

Impact evaluation and IPE integration

In this section, we harmonise the findings from the impact evaluation and the IPE. We discuss the extent to which the findings aligned with the logic model and present an interpretation of the results and key conclusions from the evaluation. Additionally, we offer recommendations for refining the logic model based on learnings from this evaluation. We link the key conclusions to the research questions and hypotheses, while also reflecting on the limitations of the evaluation.

Evidence to support the logic model

We found mixed evidence to support the logic model. On the one hand, the IPE found strong evidence that Stop and Think can be implemented as described in the logic model. On the other hand, we found mixed or weak evidence to support the short- and long-term outcomes. We will address each section of the logic model one by one.

Inputs, activities, and outputs

The IPE found strong evidence that the *inputs* and *activities* from the logic model (for example, Stop and Think software and sessions, the handbook, and the training) took place as expected. In surveys and qualitative research encounters, teachers and school leads perceived *activities*, such as training visits and ongoing support from the delivery team, to be of high quality. The software was largely regarded as user-friendly and easy to navigate, though teachers cited issues with the technology, content, and programme functionality as barriers to their engagement. On the whole, however, the relatively strong teacher engagement facilitated a smooth delivery of game sessions. It is worth noting that this was an improvement from the efficacy trial (Roy et al., 2019) where over half of the teachers reported experiencing issues with the software, which had caused delays and impeded the delivery.

One of the *activities* is the Stop and Think sessions themselves. These were to be delivered at the start of maths and/or science lessons, led by a teacher or a TA, and always completed as a whole-class activity. The IPE suggests that teachers used the programme flexibly while still mostly adhering to the prescribed model. For example, pupils interacted with the game in different ways (for example, sometimes they took turns, other times the group decided on an answer together) but 100% of respondents in the teacher post-intervention survey confirmed that sessions were completed as a whole-class activity. Interviews and observations also indicated that they were always led by a teacher or a TA.

The scheduling of Stop and Think sessions deviated somewhat from the intended model logic model. However, it should be noted that there was a discrepancy between the wording in the logic model (which described sessions taking place at the start of maths and science lessons) and the intended delivery model (which described sessions taking place at the start of maths and/or science lessons). Sixty-four percent of teachers in the post-intervention survey said that they had delivered sessions at the start of maths and science lessons, while 31% reported that the timing of their sessions has deviated from this. In interviews, school leads and teachers gave several reasons for this: for example, timetabling issues, science lessons being too infrequent, and schools having mixed year groups, which made it easier to deliver sessions during reading lessons when the other year group could read their books.

We also found strong evidence that the *outputs* of the programme took place largely as intended. The logic model describes a ten-week delivery period with 12-minute sessions three times every week. Interview and survey findings suggest that this dosage was mostly adhered to. In the post-intervention teacher survey, 72% stated that they had delivered three sessions every week and 27% said they had delivered between two and four sessions a week. Data from the software showed that 53% of classes completed all 29 sessions during the ten-week delivery period.[58] Most classes had completed at least 25 sessions after ten weeks. In the qualitative research, teachers cited timetabling and staffing issues as reasons for non-adherence.

On the whole, the IPE suggests that it is feasible to deliver the Stop and Think programme as described in the logic model. Despite the relative flexibility afforded to teachers in scheduling sessions, however, teachers were not always able to adhere to the delivery model described in the logic model of three sessions a week at the start of science and maths lessons or the intended dose of 29 sessions (not analysing the introductory session) over ten weeks. This evaluation was an effectiveness trial where the intervention was delivered without proactive support from the delivery team. Because the intervention was tested in more 'real life' circumstances and the reasons for non-adherence were practical in nature (for example, timetables, staffing), it might be difficult for teachers to adhere to the model even more strictly. However, as mentioned above, there was also a discrepancy between the precise wording of the logic model and the intended delivery model envisaged by the developers and the delivery team (delivery at start of maths and science lessons versus delivery at the start of maths and/or science lessons). Furthermore, the delivery team did not consider it critical that the sessions were delivered at the start of those lessons or precisely three times per week: the intervention developers only considered it important for the effectiveness of the intervention that children played the game regularly (approximately three sessions per week) during maths and/or science lessons, and completed as many of the sessions as possible. We have preserved the logic model in its current form as it was used as a reference point for the IPE but we recommend that it be updated, and make clear what activities and outputs are thought to be critical to achieve the outcomes and impacts.

Outcomes

The short-term outcomes in the logic model are (a) a reduction in impulsive responding to maths and science questions and (b) improved attainment on curriculum-appropriate maths and science misconceptions. These were explored through the IPE and maths and science misconception tests. The long-term outcome of improvement in attainment in maths and science academic achievement tests (c) was explored through the impact evaluation.

The IPE found mixed evidence on whether using Stop and Think was perceived to cause a reduction in pupils' impulsive responding (a). It is important to note here we were not able to collect strong evidence against this short-term outcome. We were not able to reliably measure reduction in impulsivity in pupils. Instead, at endline we collected retroactive feedback from teachers and pupils about their perceptions of pupils' impulsiveness (that is, not impulsivity itself). In focus groups, pupil views were mixed between those who had seen a positive effect in their impulsive responding and those who had not. In the post-intervention survey, only around one in four teachers thought Stop and Think had been impactful in this regard. In interviews and focus groups, teachers reasoned that while having Stop and Think as a standalone activity may have helped to reinforce the message of thinking before responding, factors such as the academic ability of the pupil and the type of question being asked still made a difference to impulsive responding. There was also a view among both teachers and pupils that the effect was limited to when pupils were playing the game and did not translate to other instances or make a lasting impact. In the qualitative research, we also heard from teachers and pupils who thought the maths items did not require any consideration time because the questions were too easy.

[58] There are 30 sessions in total but the very first session was the training session with the BIT delivery team.

We are limited in our inferences on the impact of Stop and Think on (b) improved attainment on curriculum-appropriate maths and science misconceptions. This is because the tests devised to measure this short-term outcome did not reliably measure the underlying construct of common misconceptions in those subjects. Using these tests, the impact evaluation found weak evidence of a very small and positive effect on pupils' attainment on curriculum-appropriate maths and science misconceptions. In other words, we did not find any evidence that pupils who received Stop and Think were, on average, less likely to fall into maths and science misconceptions. This question was not explored in the IPE.

For the long-term outcome (c), the impact evaluation found differential results for maths and science. For *maths*, we found that FSM-eligible pupils made no additional months' progress compared to pupils receiving teaching as usual. Regarding all pupils (both FSM-eligible and non-FSM-eligible), we found no evidence of any additional months' progress in their maths attainment compared to those receiving teaching as usual. For *science*, we found a positive and statistically significant effect on pupils' science attainment. All pupils receiving the intervention made the equivalent of two additional months' progress in science compared to pupils receiving teaching as usual. FSM-eligible pupils made one additional month's progress in their science attainment compared to FSM-eligible pupils receiving teaching as usual. Furthermore, the study was not powered enough to detect one month's progress for FSM pupils in science, and the subgroup analysis found no evidence for a statistically significant differential effect on science attainment for FSM-eligible pupils compared to their non-FSM-eligible peers.

We conducted additional mediation analysis on the primary analysis sample (pupils eligible for FSM who completed maths tests at endline). This analysis looked at the relationship between maths misconception test scores and maths progress test scores. It found that the former were predictive of pupils' maths attainment. However, we found no evidence that Stop and Think was effective in influencing these pupils' maths attainment either directly or indirectly through making them less susceptible to common misconceptions in maths. We did not conduct mediation analysis on the equivalent sample for science.

The IPE helps explain the differential impacts between maths and science. A key finding from the IPE was the perceived difference in the quality and challenge of the maths and science content. Teachers found the maths content to be aimed at lower-ability pupils and they reported more outdated terminology in the content compared to the science questions, including methods that were out of step with current teaching practice and that led to confusion and 'new' misconceptions developing in pupils. In focus groups, pupils highlighted the repetitiveness of the maths questions and were of the opinion that the questions were 'boring' and 'too easy' and that they had already covered many of the topics earlier in the school year. Here, it is important to note that the delivery period was later in the school year compared to the efficacy trial, delivered from November to March; this likely meant that pupils had had more exposure to the maths items by the time they played the game between February and May. (Note that the Key Stage 2 curriculum includes more maths than science, and so a later delivery period may have been more noticeable for maths compared to science.)

By contrast, both teachers and pupils were more positive about the science items. To begin with, they were more engaged with the content because of the perceived novelty of a science computer game. In the pre-intervention survey, 42% of teachers reported having used computer games for teaching maths in the past while only 11% had used them for science. More importantly, both teachers and pupils also considered the quality of the science items to be better. Compared to maths, teachers in the qualitative research described the difficulty level of science questions as being pitched at the right level for all pupils, not just those with lower academic ability. Pupils similarly reported that the science items were 'more fun' with more surprising answers, and that they learnt more from them compared to maths.

Furthermore, the IPE found some evidence that differences between usual practice science and maths lessons and Stop and Think were modest. However, it is important to remind the reader that the programme aims to improve inhibitory control skills in children, not increase their science and maths subject knowledge. It may have been difficult for teachers and pupils to view the programme through this lens, which is why their responses reflect on the differences they perceived in the structure and content of the programme compared to usual practice lessons. For this reason, these findings on usual practice should be treated with caution. In post-intervention interviews, teachers pointed out similarities in the content, structure, and visual style of Stop and Think and their usual lessons. They also discussed already being aware that addressing misconceptions and giving pupils time to think through their answers was beneficial. In the post-intervention survey, only about a third (31%) of teachers reported that Stop and Think was different to usual maths lessons while around half (55%) reported the same for science. Teachers and pupils agreed that the maths content was highly similar to what had already been covered in class.

Impacts

Regarding the impact 'reduced attainment gap in maths and science between advantaged and disadvantaged pupils', we conducted analysis exploring maths and science attainment gap between FSM-eligible pupils in the treatment group and their peers in the treatment group. We found that those in the treatment group did not make any significant additional months' progress in their maths and science attainment compared to their peers in the treatment group.

The other impact in the logic model is about 'improved science and maths attainment'. This falls outside the remit of our evaluation as the endline data collection took place at the end of the intervention delivery and as such we did not collect data about the pupils' longer-term attainment in maths or science.

Contextual moderating factors

The logic model highlights (a) socio-economic status and (b) compliance and dosage as the contextual moderating factors that can impact the results.

The impact evaluation and the IPE did not find evidence that socio-economic status acts as a moderator for the effect of Stop and Think on the long-term outcome of improved attainment on maths and science academic achievement tests. The impact evaluation did not find evidence supporting additional progress in the maths or science attainment of pupils who are eligible for FSM compared to pupils eligible for FSM receiving teaching as usual. These results were reflected in the IPE. In the qualitative research encounters and the post-intervention survey, teachers did not believe that pupils with a lower socio-economic status benefitted from the programme more than others. They were more likely to report that all pupils benefitted from the programme equally or that, interestingly, Stop and Think particularly benefitted SEND and EAL pupils. These findings suggest that other moderating factors may be at play, not socio-economic status, or that the programme theory is lacking in other ways.

We also did not find evidence that dosage or compliance predicted outcomes. The CACE analysis found no evidence that increased compliance led to better maths attainment for FSM pupils (our primary outcome sample). We did not carry out CACE analysis for the whole sample of pupils who took maths tests, or for the sample of pupils who took science tests. The IPE found that dosage was mostly adhered to.

Interpretation

Existing evidence suggests that pupils must be able to inhibit prior contradictory knowledge and misconceptions in order to effectively learn new concepts in maths and science. The previous efficacy trial (Roy et al., 2019) for Stop and Think found that pupils in the intervention group made the equivalent of one additional month's progress in maths and two in science, on average, compared to pupils in the control group. However, the maths result was not statistically significant. It also did not find an impact on the secondary outcome of inhibitory control, though it is worth stating that the post-intervention test used for the efficacy trial focused on speed of responding in pupils and was therefore contradictory to the aim of the programme to increase inhibitory control.

The results of our effectiveness trial are in line with those of the previous trial. Despite high compliance and adherence to dosage, we did not find an effect on our primary outcome measure of improved attainment in *maths* among pupils eligible for FSM. We also found no evidence of additional months' progress in maths for all pupils. However, Stop and Think did have a positive and statistically significant effect on pupils' attainment in *science*, our secondary outcome. All pupils receiving Stop and Think and allocated to science testing made the equivalent of two additional months' progress in science, on average, compared to pupils receiving teaching as usual. The study was not powered enough to detect one month's progress for FSM pupils in science and the subgroup analysis found no evidence for a statistically significant differential effect on science attainment for FSM-eligible pupils compared to their non-FSM-eligible peers.

The IPE findings also provide context to the differing impacts in maths and science. Teachers and pupils perceived the maths items to be of lower quality and 'too easy', perhaps partly because the delivery period was later in the school year compared to the efficacy trial. Teachers also found more outdated pedagogy in the maths items, which they worried had led to confusion and 'new misconceptions' developing in pupils. By contrast, teachers and pupils perceived the science items to be of higher quality and pitched at the right level, with the content more novel and 'surprising' to pupils. Higher engagement with the science items was partly driven by the relative novelty of a science computer game.

Regarding the logic model, we were not able to collect reliable evidence against the intended short-term outcomes of (a) reduced impulsive responding on maths and science questions and (b) improved attainment on curriculum-appropriate maths and science misconceptions. This means that our ability to interpret whether good adherence to the inputs, outputs, and activities led to the intended outcomes is limited. Similarly, the extent to which the short-term outcomes lead to the long-term outcomes for maths and science (c).

Next, we want to consider why the intervention appears to lead to differential outcomes for maths and science. Here, it is important to note that Stop and Think does not aim to improve attainment by adding to pupils' *knowledge* of science and maths; instead, the aim is to improve pupils' *inhibitory control*. According to the programme theory, exposure to the Stop and Think game, which has equal amounts of science and maths, leads to higher attainment in both subjects. This is because the game aims to improve inhibitory control in pupils, making them less susceptible to misconceptions and more adept at learning counterintuitive concepts, therefore improving their attainment in science and maths attainment tests. For this reason, the logic model does not have separate logic chains for maths and science, that is, one logic chain from the maths items in the Stop and Think game to improved maths attainment and a separate chain from the science items to improved science attainment. Despite this, our evaluation found a small positive effect on science but not on maths.

Our evaluation findings suggest that if the content of the programme is too familiar to pupils, they are not able to effectively practice their inhibitory control skills. In relation to the maths questions, both teachers and pupils reported that the questions being too easy meant that pupils did not need to reflect on their answer or 'stop and think'; that the thought process between seeing a question and retrieving the answer was near automatic. While schools on the whole completed the right number of sessions, some sessions were more effective than others and pupils, therefore, did not get to practise their 'stop and think' skills enough times over the course of the ten weeks. This helps explain the low observed impact of the intervention on pupil attainment, particularly in maths.

These findings suggest that the programme could be significantly improved by making the maths content more challenging for pupils in this age group so that the questions act as effective exercises in inhibitory control; in particular, the maths items are currently too easy for pupils. While Stop and Think is not aiming to add to pupils' knowledge of maths and science, there is an important balance to be struck in developing items that are close enough to pupils' existing knowledge about the world so that they can feasibly work out the answer, while at the same time offering stretch, challenge, and novelty to give them a chance to practice their inhibitory control skills.

Here, it is worth reflecting on the context in schools and the disproportionate amount of maths content in the Key Stage 2 curriculum compared to science. It is possible that Key Stage 2 pupils are therefore closer to their optimal abilities in maths compared to science and therefore have less scope for improvement, that is, we are witnessing ceiling effects.

Regarding the underlying programme theory, it is unclear why both the efficacy and effectiveness trials found differential impacts for science and maths when the logic model does not include separate logic chains between subject-specific content and subject-specific attainment. It appears that that links between subject-specific content and subject-specific attainment is stronger than expected. This could be because of the ceiling effects described above or because the nature of common misconceptions in science is fundamentally different to those in maths. Lastly, we found no evidence to support the theory that compliance or socio-economic status act as moderators for the intended outcomes. In addition, no dosage effect was detected within the dosages delivered. We recommend that the logic model is updated and developed further to address these gaps and inconsistencies.

Outside the logic model, the IPE found that the programme led to teachers reflecting more on their practice, particularly in science. Another significant unintended outcome was improved classroom culture. Teachers described the programme as being accessible for all because it allowed anyone to answer and because it did not matter if they got the answer wrong. Stop and Think also reduced stigma about different pupils needing different amounts of processing time, including classmates with SEND and EAL. While this is an interesting finding, it sits outside the main mechanism by which change occurs in the logic model and so we do not suggest amending the logic model on this basis.

Finally, the cost evaluation estimates that Stop and Think is a low-cost intervention. Overall, the per-pupil-per-year cost of delivery, over and above usual teaching practice, is £42.71 in total over three years, with an average of £14.24 per year. Note that the intervention was delivered by teachers. Costs could, potentially, be lower if the programme was to be delivered by TAs.

Limitations and lessons learned

While we are confident in the findings from this evaluation, there are limitations to note within the methods. In this section, we discuss limitations to the evaluation such as attrition, power, suitability of outcome measures, and missing data in different elements of the evaluation. We also reflect on lessons learnt.

Overall, school-level attrition from condition allocation to analysis was 3.5%.[59] Pupil attrition from condition allocation to analysis was 18.2% in the treatment group and 17.6% in the control group. Attrition was higher among FSM-eligible pupils who took the maths tests at endline, that is, the group who formed the sample for our primary impact evaluation analysis. The attrition level from NPD consent to analysis was 22.3% in the treatment group and 22.5% in the control group. This pupil-level attrition was exacerbated by the missing information on the NPD on pupils' FSM eligibility status.

School-level attrition was comparatively low. We also had more pupils eligible for FSM in our sample than anticipated during the design stage. This translated to a study powered to yield an MDES of 0.15 for the primary analysis.

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 were not available. The study, therefore, used KS1 maths scores for pupils in Year 3 and the EYFSP point scores for pupils in Year 5 as measures of prior attainment.

Our understanding of the programme's impact on the two short-term outcomes is limited. This is due to the inherent difficulty in measuring the intermediate aim of improved inhibitory control. The efficacy trial (Roy et al., 2019) used a measurement on inhibitory control that asked pupils to respond to as many questions as possible within a set time limit. Because the focus on speed of response may have discouraged pupils to 'stop and think', we decided against using the same instrument. Instead, we explored *perceived* effects on reduced impulsive responding through the IPE These findings are limited because they do not capture changes in impulsivity itself, only retroactive perceptions of how pupils' impulsiveness changed. We also developed new measurements to test the programme's impact on misconceptions, however, these constructs did not achieve reliability scores acceptable by convention. We did not investigate this short-term outcome in the IPE.

Our IPE findings on teachers' pre-existing knowledge of counterintuitive concepts and inhibitory control is limited by the fact that teachers were already part of the trial when we asked about their awareness and application of these concepts. We asked teachers about this in the pre-intervention survey (by which point they would have been recruited for the trial and been given information about the programme aims and, in many cases, already completed the Stop and Think training) and in interviews during programme delivery (by which point trial teachers would have encountered at least some of the content). This limitation in measuring teachers' pre-existing knowledge meant that we could not objectively assess the difference the programme made in this regard.

Finally, while we heard a range of views in the IPE, we are limited to the perspectives of those who took part in these elements of the evaluation. There is a risk that the IPE survey, interview, and focus group data from teachers, pupils, and school leads is skewed by the responses coming from individuals with greater interest and awareness of Stop and Think than those who did not respond to the survey or participate in interviews or focus groups. It is possible that other perspectives, which could have influenced findings, are not captured in the IPE results included in this evaluation. In particular, the responses rates to the surveys from control group teachers were lower compared to the response rates of intervention group teachers. We are continuing to reflect on ways to minimise the data collection burden on evaluation participants, particularly busy professionals with limited time, to increase engagement in the IPE elements of evaluation.

Future research and publications

The IPE found that Stop and Think was delivered with fidelity and that teachers found it user-friendly to deliver the programme as it was designed. Since the science element of the programme was deemed as 'novel' and had a small positive and statistically significant impact, another study could explore science attainment as its primary outcome to

[59] Due to the ONS SRS disclosure guidance, we are not able to share school-level attrition from condition allocation to NPD consent, from NPD consent to analysis, and condition allocation to analysis to prevent primary and secondary statistical disclosure. SRS Statistical Disclosure Control Policy for DfE data (Office for National Statistics, 2024). Available at https://assets.publishing.service.gov.uk/media/660d8798758315001a4a49d2/DfE_ONS_statistical_disclosure_control_policy.pdf

understand better the mechanisms at play. It could also be worthwhile to explore alternative outcome measurements for pupils that can accurately measure the two short-term outcomes on the logic model.

There are no planned additional publications for this trial.

References

Archer Ker, L. and Tomei, A. (2013) 'What Influences Participation in Science and Mathematics?: A Briefing Paper from the Targeted Initiative on Science and Mathematics Education (TISME)': https://kclpure.kcl.ac.uk/portal/files/64435093/TISME_briefing_paper_March_2013.pdf

Babai, R., Shalev, E. and Stavy, R. (2015) 'A Warning Intervention Improves Students' Ability to Overcome Intuitive Interference', *ZDM Mathematics Education* 47, pp. 735–745: https://link.springer.com/article/10.1007/s11858-015-0670-y#article-info

Borenstein, M., Hedges, L. V., Higgins, J. P. and Rothstein, H. R. (2009) *Introduction to Meta-Analysis*, John Wiley.

Carlson, S. M., Moses, L. J. and Breton, C. (2002) 'How Specific is the Relation Between Executive Function and Theory of Mind? Contributions of Inhibitory Control and Working Memory', *Infant and Child Development*, 11 (2), pp. 73–92: https://psycnet.apa.org/record/2002-17339-002

Cortina, J. M. (1993) 'What is Coefficient Alpha? An Examination of Theory and Applications', *Applied Psychology*, 78 (1), p. 98.

Department for Science, Innovation and Technology (2024) 'About us - Department for Science, Innovation and Technology - GOV.UK (www.gov.uk)

DfE (2020) 'Multi-Million Government Investment in the Future of UK Science': https://www.gov.uk/government/news/multi-million-government-investment-in-the-future-of-uk-science

Durham University (2020) 'Re-Analysis: Stop and Think: Learning Counterintuitive Concepts', (unpublished manuscript).

Dracup, T. (2014) 'The Politics of Setting': https://giftedphoenix.wordpress.com/2014/11/12/the-politics-of-setting/

EEF (2022) 'Statistical Analysis Guidance for EEF Evaluations': https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1728854692

EEF (2019) 'Classification of the Security of Findings from EEF Evaluations': https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/Classifying_the_security_of_EEF_findings_2019.pdf

Flora (2020) 'Your Coefficient Alpha Is Probably Wrong, but Which Coefficient Omega is Right? A Tutorial on Using R to Obtain Better Reliability Estimates', *Advances in Methods and Practices in Psychological Science*, 3 (4), pp. 484–501. DOI: 10.1177/2515245920951747

Gauthier, A., Porayska-Pomsta, K., Mayer, S., Dumonteil, I., Farran, E., Bell, D., Mareschal, D. and the UnLocke Team (2022) 'Redesigning Learning Games for Different Learning Contexts: Applying a Serious Game Design Framework to Redesign Stop & Think', *Child-Computer Interaction*, 33, 100503: doi.org/10.1016/j.ijcci.2022.100503

Kuczera, M., Field, S. and Windisch, H. C. (2016) 'Building Skills for All: A Review of England': http://www.oecd.org/education/skills-beyond-school/building-skills-for-all-review-of-england.pdf

Luedtke, O., Robitzsch, A. and Grund, S. (2017) Multiple Imputation of Missing Data in Multilevel Designs: A Comparison of Different Strategies', *Psychological Methods*, 22 (1), pp. 141–165. https://doi.org/10.1037/met0000096

McKaskill, M., Schei, T., and Fugard, A. (2025) **'**Developing Tests for Common Misconceptions in Key Stage 2 Maths and Science: Technical Report'

Morgan, R., Kirby, C. and Stamenkovic, A. (2016) 'The UK STEM Education Landscape: A Report for the Lloyd's Register Foundation from the Royal Academy of Engineering Education and Skills Committee': https://www.raeng.org.uk/publications/reports/uk-stem-education-landscape

NFER (2016) 'Protocol for the Evaluation of Counterintuitive Concepts Intervention': https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/Neuroscience_-_Counterintuitive_Concepts_Protocol_AMENDED.pdf?v=1723788250

OECD (2017) 'Survey for Adult Skills (PIACC)': https://www.oecd.org/skills/piaac/

OECD (2022) 'PISA 2022: PISA Results in Focus': https://www.oecd.org/publication/pisa-2022-results/index

Roy, P., Rutt, S., Easton, C., Sims, D., Bradshaw, S., McNamara, S. (2019) 'Stop and Think: Learning Counterintuitive Concepts, Evaluation Report': chromeextension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.nfer.ac.uk/media/uqapkjif/learning_counterintuitive_concepts_evaluation_report_-final.pdfUK Commission for Employment and Skills (2022) 'Skills for the Workplace: Employer Perspectives': https://www.gov.uk/government/publications/skills-for-the-workforce-employer-perspectives

van Buuren, S. and Groothuis-Oudshoorn, K. (2011) 'Multivariate Imputation by Chained Equations', *Statistical Software*, 45 (3), pp. 1–67. https://doi.org/10.1177/0962280206074463

van Buuren, S. and Oudshoorn, C. G. M. (2000) 'Multivariate Imputation by Chained Equations' (MICE V1.0 user mannual), TNO Prevention and Health, Public Health.

Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A. and Kyrianakis, G. (2018) 'The Recruitment of Shifting and Inhibition in on-Line Science and Mathematics Tasks', *Cognitive Science* 42 (6), pp. 1860–886. https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12624?af=R

Wilkinson, H. R. , Smid, C., Morris, S., Farran, E. K., Dumontheil, I., Mayer, S., Tolmie, A., Bell, D., Porayska-Pomsta, K., Holmes, W., Mareschal, D., Thomas, M. S. C. and the UnLocke Team (2019) 'Domain-Specific Inhibitory Control Training to Improve Children's Learning of Counterintuitive Concepts in Mathematics and Science', *Cognitive Enhancement*: doi.org/10.1007/s41465-019-00161-4

Woods, A. D., Gerasimova, D., van Dusen, B., Nissen, J., Bainter, S., Uzdavines, A., Davis-Kean, P. E., Halvorson, M., King, K. M., Logan, J. A. R., Xu, M., Vasilev, M. R., Clay, J. M., Moreau, D., Joyal-Desmarais, K., Cruz, R. A., Brown, D. M. Y., Schmidt, K. and Elsherif, M. M. (2024) 'Best Practices for Addressing Missing Data Through Multiple Imputation', *Infant and Child Development*, 33 (1), e2407. https://doi.org/10.1002/icd.2407

Appendix A: EEF cost rating

Table 46. Cost Rating

Cost rating	Description
£ £ £ £ £	*Very low:* less than £80 per pupil per year.
£ £ £ £ £	*Low:* up to about £200 per pupil per year.
£ £ £ £ £	*Moderate:* up to about £700 per pupil per year.
£ £ £ £ £	*High:* up to £1,200 per pupil per year.
£ £ £ £ £	*Very high:* over £1,200 per pupil per year.

Appendix B: Security classification of trial findings

OUTCOME: *Age-standardised* GL Progress Tests in Maths (FSM pupils only)

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	**MDES**	**Attrition**			Adjustment for threats to internal validity **[N/A]**		
5 🔒	Randomised design	<= 0.2	0-10%					
4 🔒	Design for comparison that considers some type of selection on unobservable characteristics (e.g. RDD, Diff-in-Diffs, Matched Diff-in-Diffs)	0.21 - 0.29	11-20%					
3 🔒	Design for comparison that considers selection on all relevant observable confounders (e.g. Matching or Regression Analysis with variables descriptive of the selection mechanism)	0.30 - 0.39	21-30%	3				3
2 🔒	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1 🔒	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0 🔒	No comparator	>=0.6	>50%					

Threats to validity	Risk rating	Comments
Threat 1: Confounding	Low	Randomised design, and good balance between the groups minimised the risk of confounding.
Threat 2: Concurrent Interventions	Low	Explored within the IPE. No obvious use of concurrent interventions in schools – particularly with science in primary schools as reported by school leads. However, there is some evidence of varying awareness among teachers of 'stop and think' as an approach. Usual practice differences between treatment and control in science and maths lessons were modest, although there was a poor response rate amongst control teachers.
Threat 3: Experimental effects	Low	Although two schools were misallocated, an examination of the results found no evidence that this had an impact.
Threat 4: Implementation fidelity	Low	Intervention was delivered with fidelity.
Threat 5: Missing Data	Moderate	No obvious differential dropout amongst pupils, and the school level dropout is low at 3.5%. However, missing data is higher than ideal at 18.3 and 17.6% for treatment and control group. Analysis does not reveal any large impact of missing data
Threat 6: Measurement of Outcomes	Low	The GL PTM assessment post-test measure was a well validated and appropriate commercial test, conducted by independent administrators. Possible ceiling effect on PTM tests, but all relatively minor.
Threat 7: Selective reporting	Low	No evidence of selective reporting (but note probable low response to teacher surveys)

- **Initial padlock score:** 3 padlocks – two-arm cluster randomised trial; MDES of .14 at randomisation. 22.48% pupil attrition led to a loss of two padlocks.
- **Reason for adjustment for threats to validity**: None
- FINAL PADLOCK SCORE: 3

Appendix C: Changes since the previous evaluation

Table 47. Changes since the previous evaluation

Feature		Efficacy to effectiveness stage
Intervention	**Intervention content**	Content: ● Easier questions will be dropped, and more difficult questions added (e.g. Year 6 maths), based on feedback during the efficacy trial. ● New illustration work will be extended to Year 3 and 5 tasks. ● Content to be reviewed and updated where needed to ensure accuracy. ● Teachers will be able to choose from weekly themes (e.g. fractions, animals – using a teacher dashboard) in response to teacher feedback during the efficacy trial. ● Teachers will still be able to opt for random allocation of themes, which was the default during the efficacy trial. ● Optional motivational elements will be added: a leader board (where classes can see how many sessions they have completed, how they stand in relation to other schools) and virtual coins (for each session completed, which can be used to buy items to improve an animal avatar). The aim is to build motivation for classrooms to engage with Stop and Think activities and foster a collaborative atmosphere between pupils. ● Birkbeck will be conducting the following activities during the development phase: ○ A design workshop with game designers, teachers and the unLocke team to explore design ideas with respect to the above design goals. ○ A focus group with children to appraise the design of the software in terms of its appeal and motivation to engage with specific tasks. ○ A small-scale validation study in schools to identify any remaining software issues. Software: ● The application will be restructured to be more robust, reliable and flexible, including: ○ A fundamental restructuring of the flow of the application to use co-routines instead of statically-timed steps, to which between phases of interaction ○ UI projections instead of static images to display task previous and correct answer views ○ A home-screen where the teacher can access new sessions and revisit past sessions. ● The database will be rebuilt to facilitate new teacher registration and data-saving.
	Delivery model	● During the efficacy trial, Stop and Think assistants kept in regular proactive contact with schools, including following up with the programme when it was not being used regularly. ● This support will not be provided during the effectiveness trial, based on recommendations from NatCen to replicate more 'real world' conditions. ● Teachers will still meet an assistant in person during the initial information session, and will be able to consult an FAQ document, briefing video and handbook. Teachers will also be able to reach the BIT delivery team via email and telephone during the intervention period. ● Robust web-based teacher-training materials will be created (covering how to run the module effectively, plus troubleshooting). ● Reminders are now built into the system. If a teacher has started at least one Stop and Think session since Monday at 12am, a reminder email is sent on Friday morning at 7am. If a teacher only has one session to complete, the email will remind them to complete that session on that day. If a teacher has two sessions to complete, the email will remind them to complete one session on that day and recommend completing four sessions the following week. No reminder email will be sent if a teacher has completed three sessions that week (per recommendation).
	Intervention duration	No change.

Evaluati on	**Eligibility criteria**	Schools which participated in the Stop and Think efficacy trial, the piloting work to refine the software for the Stop and Think effectiveness trial or piloting work for the misconceptions test will not be eligible for the trial.
	Level of randomisation	No change.
	Outcomes and baseline	● Change from co-primary outcomes (maths and science attainment) for all pupils to a single primary outcome (maths attainment) for pupils eligible for FSM as the primary population of interest. ● No change in outcome measures. ● No change in baseline measure used for Year 5 pupils (Early Years Foundation Stage Profile). ● KS1 test scores now used as baseline measures for Year 3 pupils. ● New test developed for intermediate outcome to measure common maths and science misconceptions.
	Control condition	● The effectiveness trial is a two-arm cluster randomised trial. Therefore, no control-plus condition (implementing a social skills programme called See+) will be used and only a business-as-usual control condition will be used in addition to the intervention.

Appendix D: Distribution of outcomes at endline for both groups

Figure 21. Distribution of age-standardised GL PTM score band for primary analysis

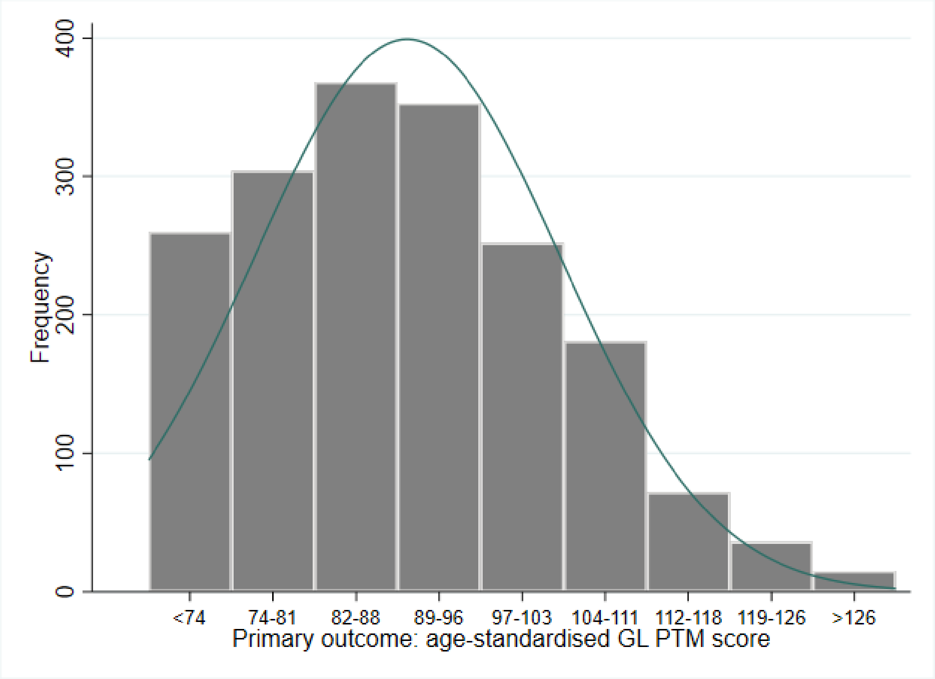


Figure 22. Distribution of age-standardised GL PTM score band at endline – All maths pupils

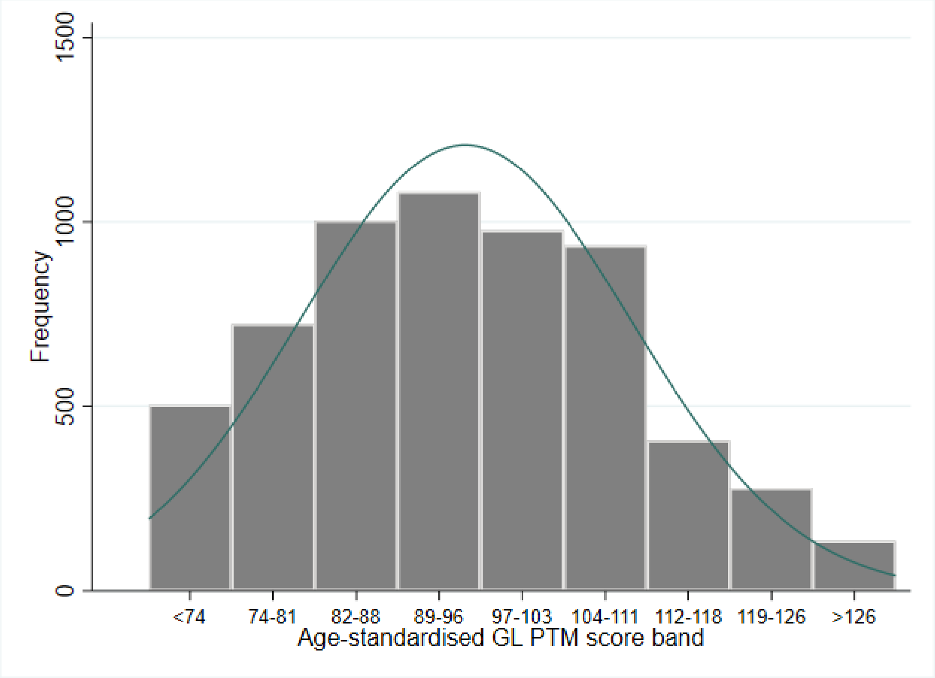


Figure 23. Distribution of age-standardised GL PTS score band at endline – all science pupils

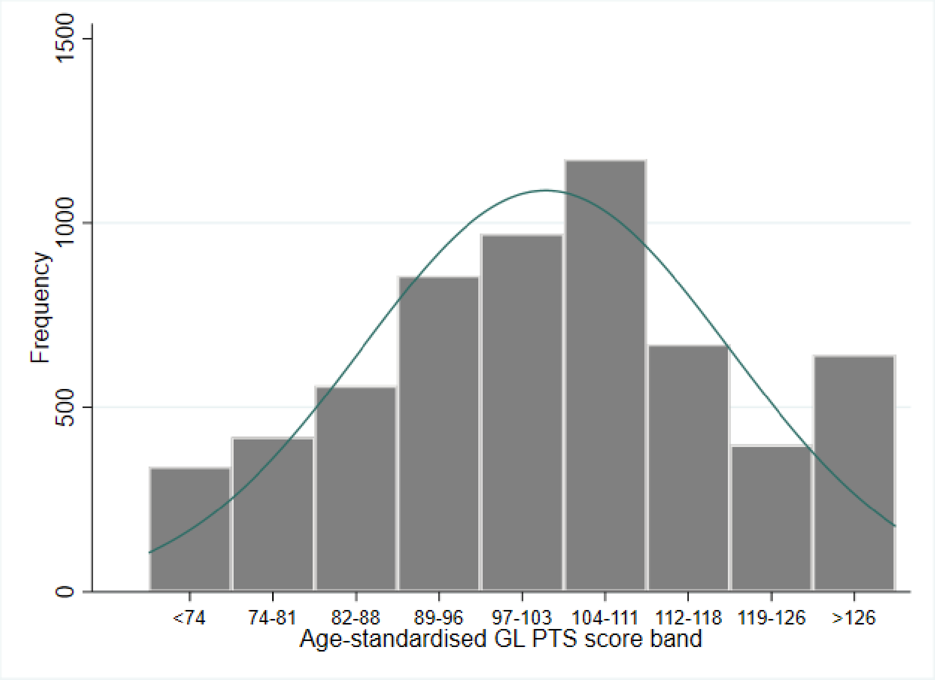


Figure 24. Distribution of age-standardised GL PTS score band at endline – science pupils eligible for FSM

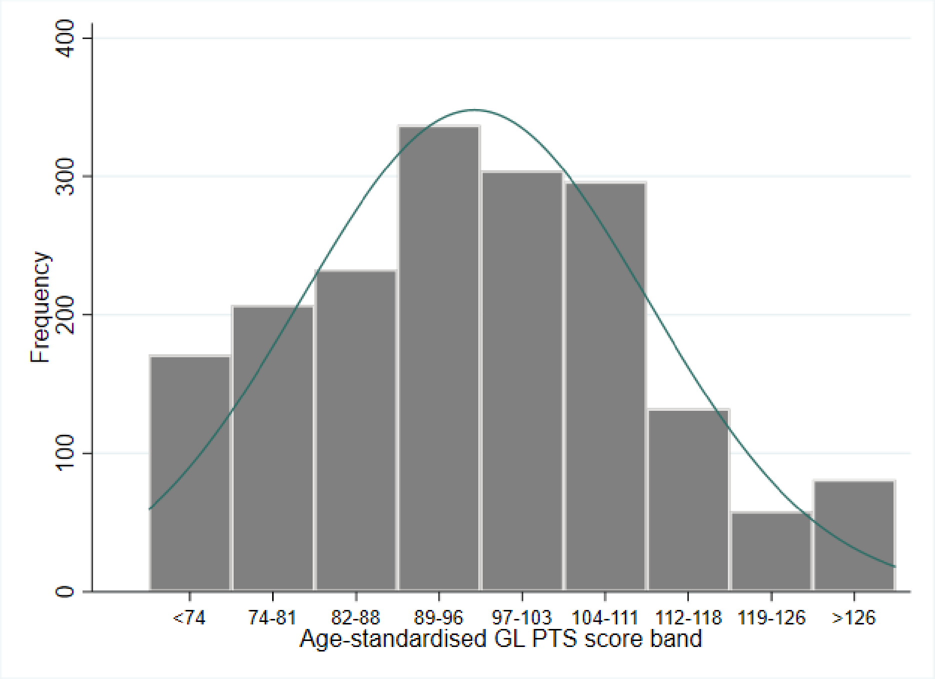


Figure 25. Distribution of validated raw score of misconception in maths – all maths pupils

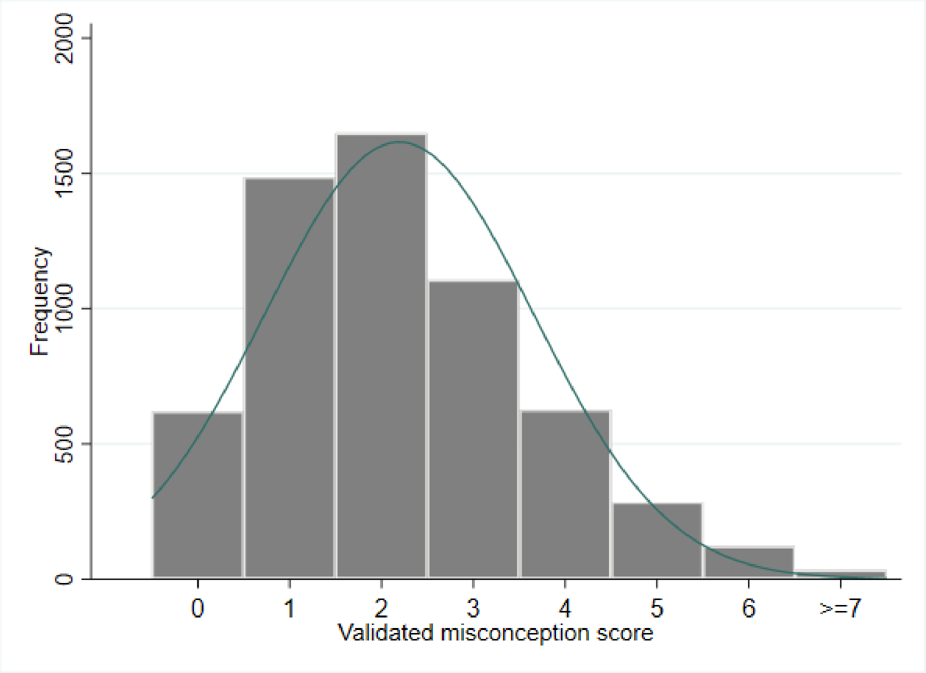
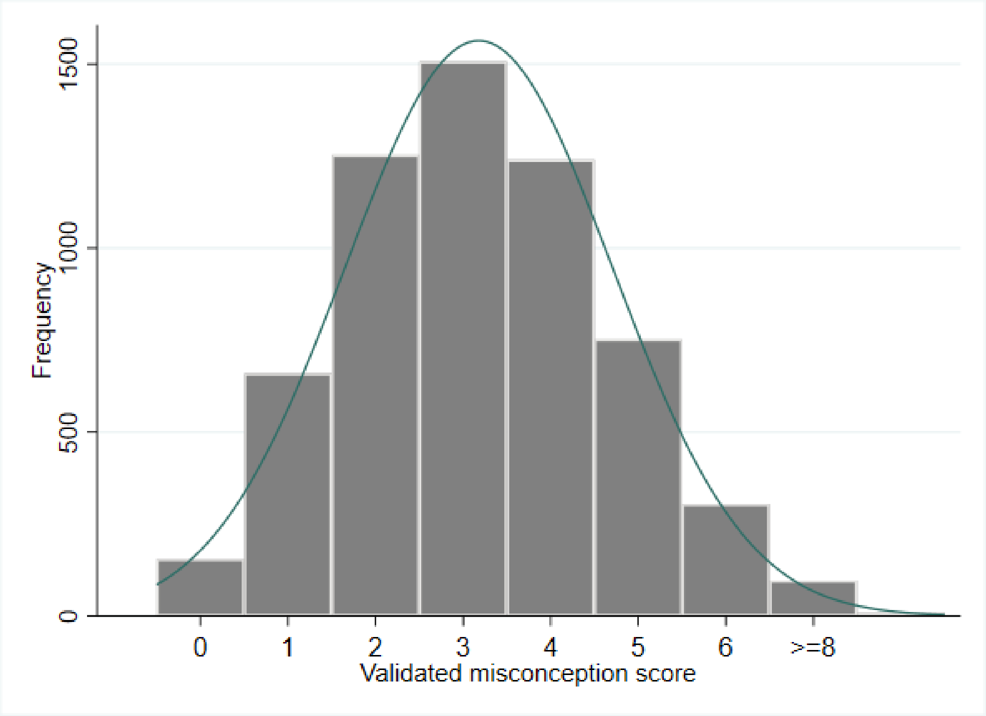


Figure 26. Distribution of validate raw score of misconception in science – all science pupils



Appendix F: Missing data analysis

Table 48. Determinants of data missingness and missing outcome measure – PTM

	(1)	(2)
VARIABLES	All Maths	All Maths
class_form	-0.108	
	(0.202)	
schoolFSMpc	0.000509	
	(0.00618)	
2o.EVERFSM_6_P_SPR23_m (reference)	-	
baseline_score_cor	-0.587***	
	(0.0371)	
o.baseline_score_m	-	
1.treat	-0.0395	
	(0.0758)	
2.strata_rantreat	-0.349	
	(0.275)	
3.strata_rantreat	-0.395	
	(0.427)	
4.strata_rantreat	-0.508	
	(0.322)	
5.strata_rantreat	-0.185	
	(0.238)	
6.strata_rantreat	-0.107	
	(0.388)	
7.strata_rantreat	-0.323	
	(0.372)	
8.strata_rantreat	-0.180	
	(0.273)	
9o.strata_rantreat (reference)	-	
5.year_group	-0.0708	
	(0.0758)	
var(_cons[SchoolSerial])		0.493***
		(0.0894)
1.EVERFSM_6_P_SPR23_m	0.490***	
	(0.0844)	
Constant	-1.856***	
	(0.329)	
Observations	6,968	6,968
Number of groups	169	169

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Coefficients from the melogit regression of a binary variable indicating whether the outcome and covariates of interest were missing. All specifications also include stratification block fixed-effects. Baseline score was imputed with the mean if missing

Appendix G: Imputation analysis

Table 49. Determinants of missingness in outcome measure of PTM informing the MI process

	(1)	(2)
VARIABLES	All Maths	All Maths
class_form	-0.127	
	(0.196)	
SchoolFSMpc	0.000283	
	(0.00600)	
2o.EVERFSM_6_P_SPR23_m (reference)	-	
baseline_score_cor	-0.585***	
	(0.0369)	
baseline_score_m	0.998***	
	(0.151)	
1.treat	-0.0437	
	(0.0730)	
2.strata_rantreat	-0.399	
	(0.267)	
3.strata_rantreat	-0.376	
	(0.413)	
4.strata_rantreat	-0.599*	
	(0.312)	
5.strata_rantreat	-0.222	
	(0.230)	
6.strata_rantreat	-0.0749	
	(0.377)	
7.strata_rantreat	-0.379	
	(0.360)	
8.strata_rantreat	-0.223	
	(0.264)	

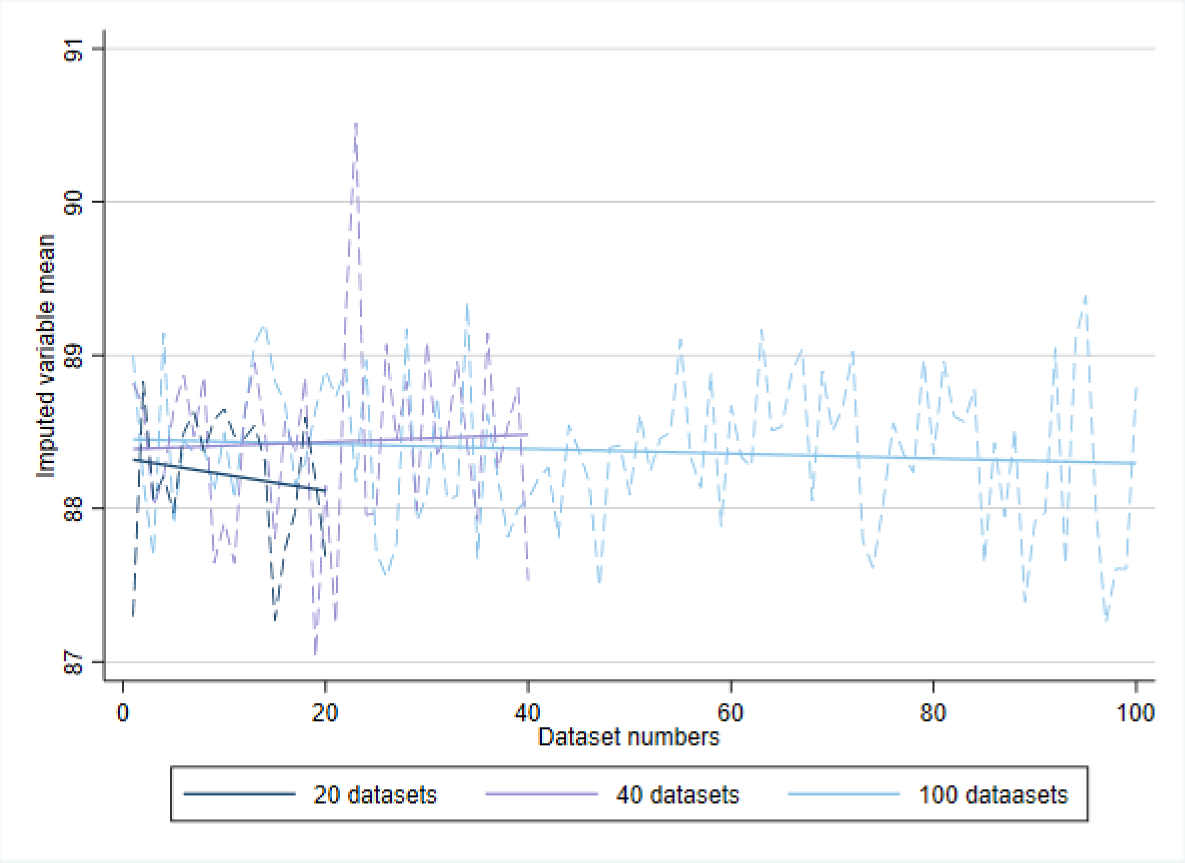
9o.strata_rantreat (reference)	-	
5.year_group_Y5	-0.111	
	(0.0743)	
var(_cons[SchoolSerial])		0.466***
		(0.0844)
1.EVERFSM_6_P_SPR23_m	0.503***	
	(0.0812)	
Constant	-1.756***	
	(0.317)	
Observations	7,265	7,265
Number of groups	169	169

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Coefficients from the melogit regression of a binary variable indicating whether the outcome of interest was missing. All specifications also include stratification block fixed-effects. Baseline score was imputed with the mean if missing

Figure 27. Age-standardised PTM MI convergence



Appendix H: Additional sensitivity analysis

Sensitivity analysis – FSM maths

Table 50. Results of sensitivity analysis for maths pupils eligible for FSM

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Age-standardised PTM score – Year 3	0.69 (-1.05, 2.42)	0.95 (-0.79, 2.69)	452(107)	126.61	447(113)	136.68	131.61
Age-standardised PTM score – Year 5	0.33 (-0.94, 1.6)	-0.01 (-2.17, 2.15)	491(165)	204.21	451(149)	193.39	199.05

Table 51. Results of sensitivity analysis for maths pupils eligible for FSM – effect size estimation

Outcome	Unadjusted means				Effect size			Combined Effect Size
	Intervention group		Control group					
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value	
Age-standardised PTM score – Year 3	452(107)	89.83 (88.63, 91.03)	447(113)	89.15 (87.89, 90.41)	899 (452; 447)	0.07 (-.06, .2)	0.282	0.04
Age-standardised PTM score – Year 5	491(165)	89.46 (88.16, 90.77)	451(149)	89.48 (88.17, 90.78)	942 (491; 451)	0.00 (-0.13, 0.13)	0.998	

Sensitivity analysis – All maths

Table 52. Results of sensitivity analysis for all maths pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		Pooled variance
			n (missing)	variance of outcome	n (missing)	variance of outcome	
Raw PTM score – Year 3	0.28 (-0.58, 1.14)	0.95 (-0.23, 2.14)	1560(278)	145.24	1514(265)	150.52	147.84
Raw PTM score – Year 5	0.84 (-0.26, 1.93)	-0.03 (-1.85, 1.80)	1520(363)	236.53	1450(344)	227.27	232.01

Table 53. Results of sensitivity analysis for all maths pupils – effect size estimation

Outcome	Unadjusted means				Effect size			Combined Effect Size
	Intervention group		Control group					
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value	
Raw PTM score – Year 3	1560(278)	25.74 (25.15, 26.34)	1514(265)	25.46 (24.84, 26.08)	3074 (1560; 1514)	0.08 (0.01, 0.15)	0.113	0.04 (-0.01, 0.09)
Raw PTM score – Year 5	1520(363)	32.02 (31.24, 32.79)	1450(344)	31.18 (30.4, 31.96)	2970 (1520; 1450)	0.00 (-0.07, 0.07)	0.976	

Sensitivity analysis – All science

Table 54. Results of sensitivity analysis for all science pupils

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		
			n (missing)	variance of outcome	n (missing)	variance of outcome	Pooled variance
Raw PTS score – Year 3	0.96 (0.46, 1.45)	1.15 (0.45, 1.85)	1524(308)	49.38	1525(259)	47.85	48.61
Raw PTS score – Year 5	0.07 (-0.56, 0.69)	0.15 (-0.84, 1.13)	1538(348)	74.28	1446(345)	76.67	75.43

Table 55. Results of sensitivity analysis for all science pupils – effect size estimation

	Unadjusted means				Effect size			Combined Effect Size
	Intervention group		Control group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value	
Raw PTS score – Year 3	1524(308)	24.15 (23.8, 24.51)	1525(259)	23.2 (22.85, 23.54)	3049 (1524; 1525)	0.16 (0.09, 0.24)	0.001	0.09 (0.04, 0.13)
Raw PTS score – Year 5	1538(348)	29.8 (29.37, 30.23)	1446(345)	29.73 (29.28, 30.18)	2984 (1538; 1446)	0.02 (-0.05, 0.09)	0.769	

Sensitivity analysis – FSM Science

Table 56 Results of sensitivity analysis for science pupils eligible for FSM

Outcome	Unadjusted differences in means (I-C) (95% CI)	Adjusted differences in means (I-C) (95% CI)	Intervention group		Control group		
			n (missing)	variance of outcome	n (missing)	variance of outcome	Pooled variance
Raw PTS score – Year 3	.92 (0, 1.84)	1.23 (.22, 2.24)	436(122)	50.96	436(121)	45.23	48.1
Raw PTS score – Year 5	-0.51 (-1.59, 0.58)	-0.06 (-1.32, 1.20)	487(148)	70.02	460(142)	75.99	72.92

Table 57 Results of sensitivity analysis for science pupils eligible for FSM – effect size estimation

	Unadjusted means				Effect size			Combined Effect Size
	Intervention group		Control group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges g (95% CI)	p-value	
Raw PTS score – Year 3	436(122)	21.38 (20.7, 22.05)	436(121)	20.46 (19.82, 21.09)	872 (436; 436)	0.18 (0.04, 0.31)	0.017	0.08 (-0.01, 0.17)
Raw PTS score – Year 5	487(148)	26.16 (25.42, 26.91)	460(142)	26.67 (25.87, 27.47)	947 (487; 460)	-0.01 (-0.13, 0.12)	0.929	

Comparison of effect size by year group and subject between the current effectiveness trial and efficacy trial

Sample	Outcome	Effectiveness trial	Efficacy trial (Roy et al., 2019)

		n in model (intervention; control)	Hedges g (95% CI)	p-value	combined effect size Y3 and Y5 (95% CI)	n in model (intervention; control)	Hedges g (95% CI)	p-value	combined effect size Y3 and Y5 (95% CI)
FSM maths	Raw PTM8 score – Year 3	899 (452; 447)	0.07 (-0.06, 0.2)	0.282	0.04 (-0.06, 0.12)	381 (179; 202)	0.19 (-0.02, 0.40)	0.07	
	Raw PTM10 score – Year 5	942 (491; 451)	0.00 (-0.13, 0.13)	0.998		444 (246; 198)	0.16 (-0.04, 0.36)	0.11	
All maths	Raw PTM8 score – Year 3	3074 (1560; 1514)	0.08 (0.01, 0.15)	0.113	0.04(-0.01, 0.09)	1326 (647; 679)	0.03 (-0.12, 0.18)	0.67	0.09(-0.01,0.19)
	Raw PTM10 score – Year 5	2970 (1520; 1450)	0.00 (-0.07, 0.07)	0.976		1376 (696; 680)	**0.14 (-0.002, 0.28)**	**0.05**	
All science	Raw PTS8 score – Year 3	3049 (1524; 1525)	**0.16 (0.09, 0.24)**	**0.001**	**0.09 (0.04, 0.13)**	1354 (651; 703)	0.07 (-0.08, 0.22)	0.34	**0.12(0.02,0.22)**
	Raw PTS10 score – Year 5	2984 (1538; 1446)	0.02 (-0.05, 0.09)	0.769		1381 (693; 688)	**0.17 (0.03, 0.32)**	**0.02**	
FSM science	Raw PTS8 score – Year 3	872 (436; 436)	**0.18 (0.04, .31)**	**0.017**	0.08 (-0.01, 0.17)	377 (175; 202)	0.01 (-0.19, 0.20)	0.96	
	Raw PTS10 score – Year 5	947 (487; 460)	-0.01 (-0.13, 0.12)	0.929		442 (245; 197)	0.10 (-0.13, 0.33)	0.39	

Interaction term subgroup analysis – maths attainment for intervention vs control group in FSM-eligible subsample

Table 58. Subgroup analysis for maths – intervention and control group in the FSM-eligible subsample

	Unadjusted differences in means (I-C) (95% CIs)	Adjusted differences in means (I-C) (95% CIs)	Intervention group		Control group		
Outcome			n (missing)	variance of outcome	n (missing)	variance of outcome	Pooled variance
Age-standardised PTM score	.33 (-.94, 1.6)	.5 (-.86, 1.86)	943(272)	193.55	898(262)	191.32	192.46

Table 59. Subgroup analysis for maths – intervention and control group in the FSM-eligible subsample – effect size estimations

	Adjusted means				Effect size			
	Intervention group: FSM		Control group: FSM					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n FSM (intervention; control)	Raw coefficient (95% CI)	ES	p-value
Age-standardised PTM score	943(272)	92.97 (91.96, 93.97)	898(262)	92.47 (91.55, 93.39)	1841 (943; 898)	-0.25 (-1.62, 1.12)	0.04	0.720

Interaction term subgroup analysis – science attainment for intervention vs control group in FSM-eligible subsample

Table 60. Subgroup analysis for science – intervention and control group in the FSM-eligible subsample

			Intervention group		Control group		
Outcome	Unadjusted differences in means (I-C) (95% CIs)	Adjusted differences in means (I-C) (95% CIs)	n (missing)	variance of outcome	n (missing)	variance of outcome	Pooled variance
Age-standardised PTS score	0.63 (-0.84, 2.1)	1.6 (-0.12, 3.32)	923(270)	266.3	896(263)	242.89	254.77

Table 61. Subgroup analysis for science – intervention and control group in the FSM-eligible subsample – effect size estimations

	Adjusted means				Effect size			
	Intervention group: FSM		Control group: FSM					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n FSM (intervention; control)	Raw coefficient (95% CI)	ES	p-value
Age-standardised PTS score	923(270)	100.42 (99.21, 101.63)	896(263)	98.82 (97.60, 100.04)	1819 (923; 896)	-0.30 (-1.95, 1.34)	0.10	0.719

Appendix H: Trial recruitment materials

Stop and Think: Memorandum of Understanding (MoU)

If you are happy for your school to take part in the Stop and Think programme, please complete this MoU and send a scanned copy (both sides) to stopandthink@bi.team by [DATE].

Stop and Think MoU: agreement to participate in intervention and evaluation

School Leads: please read the following statements and initial the boxes if you agree with each of them.

I confirm that I have read and understood the '**Stop and Think: Learning Counterintuitive Concepts'** information sheet and have had the opportunity to ask questions. □

I understand that by agreeing to use the Stop and Think programme, my school will also be taking part in the independent evaluation conducted by NatCen Social Research (NatCen). □

I understand that the Behavioural Insights Team is responsible for delivering the project and NatCen is responsible for the evaluation. For any questions:

stopandthink@bi.team (for questions about the software or how the programme will run).

StopAndThinkEvaluation@natcen.ac.uk (for questions about the evaluation or the way we will use data). □

I understand that the evaluation includes a randomised control trial, and that Year 3 and Year 5 in my school will be randomly assigned to either use the Stop and Think programme ('treatment group') or to continue teaching as usual ('control group'). □

I understand that the Year 3 or Year 5 teachers in the treatment group will deliver 30 Stop and Think sessions in maths and science lessons over 10 weeks, each lasting 15 minutes. I understand that the school must meet the minimum technological requirements of the project (computer and a projector/ interactive whiteboard) to deliver the sessions. □

I understand that a Stop and Think 'Project Champion' from the Behavioural Insights Team (BIT) will come to the school to install the Stop and Think software and to deliver a short initial training session on how to use it. I agree to facilitate teachers in the treatment group taking part in this session. □

I agree to share the following information for all Year 3 and Year 5 pupils with NatCen: Unique Pupil Number, date of birth, full name, FSM eligibility and class. □

I understand that the developers of Stop and Think at Birkbeck University and researchers at NatCen will have access to software data that shows how classes have used the software. □

I understand that the evaluation will involve a researcher coming into the school to run maths and science tests with all Year 3 and Year 5 pupils in Summer 2023. I agree to facilitate this. □

I agree to facilitate NatCen's other evaluation activities with pupils and staff in my school. This will include teacher surveys and might include focus groups with teachers and pupils and □

observations of maths and science lessons.

I agree to circulate parent/carer withdrawal forms in September 2022 and to notify NatCen of any pupils that are withdrawn from the evaluation throughout the project. □

I have read the NatCen's Privacy Notice and understand that the NatCen research team will store information collected from staff and pupils securely and only share it with designated individuals for the purposes of the research. □

Test results will be linked with information about the pupils from the National Pupil Database (NPD). At the end of the evaluation, the data will be shared with Birkbeck University, the Department for Education, FFT Education (EEF's data processor for their archive), and the Office for National Statistics. Data may also be shared in an anonymised form with other research terms in the future. All EEF trial data is stored in the EEF data archive. The archive does not contain direct identifiers like pupil name, contact details and date of birth, but does hold a Pupil Matching Reference (PMR). The PRM is used for further matching to the NPD and other administrative datasets that may be required as part of subsequent research. We will not use pupil names or school names in any report arising from this research. □

I understand that the evaluation has been reviewed by and received ethical approval through the Research Ethics Committee (REC) at NatCen Social Research. □

I know who I can contact if I have any concerns or complaints about either the Stop and Think programme or NatCen's evaluation. □

I understand that my school's participation in the intervention and the evaluation is voluntary and that I am free to withdraw at any time, without giving any reason. □

Stop and Think: MoU signing page

Please complete Part 1 and ask your headteacher to sign Part 2.

Part 1

School name: ____

School address: ____

School postcode: ____

Estimated pupil and class details for the **2022 - 23 academic year:**

An estimate is fine if you do not have the exact information.

Number of year 3 pupils: □ Number of year 3 classes: □

Number of year 5 pupils: □ Number of year 5 classes: □

The main contact for the evaluation will be:

Name: ____

Job title: ____

Contact phone number: ____

Email: ____

The information in this MoU will be collected and stored by the Behavioural Insights Team, in accordance with its Privacy Notice. It will be shared securely with NatCen so that NatCen can organise its evaluation activities. Please see the evaluation Privacy Notice prepared by NatCen for more details on the pupil, teacher and school level data that will be collected during the evaluation.

Part 2

I am happy that my school **will take part in the Stop and Think intervention and evaluation** and agree to the conditions stated in this MoU.

Headteacher signature: ____________________________________

Headteacher name: ____________________________________

If you have any queries about this project, please contact:

stopandthink@bi.team (for questions about the software or how the programme will run); or

StopandThinkEvaluation@natcen.ac.uk (for questions about the evaluation or the way we will use data).

Stop and Think: Learning Counterintuitive Concepts

Information sheet for schools

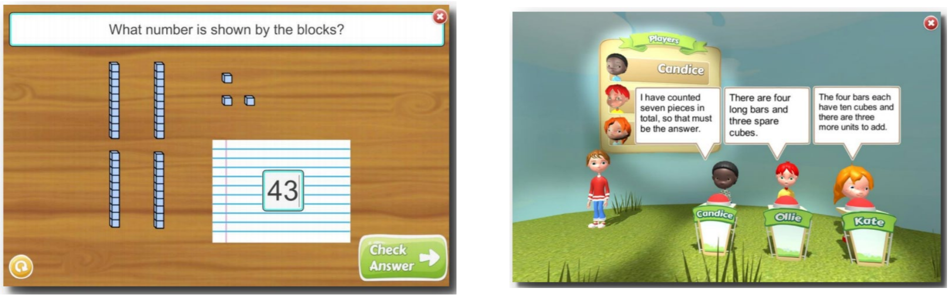
This study is a collaboration between the Education Endowment Foundation (EEF), the Behavioural Insights Team (BIT), the National Centre for Social Research (NatCen) and the Centre for Educational Neuroscience at Birkbeck, University of London.

Summary

- We invite state primary schools to take part in a project testing a computer-assisted learning programme, *Stop and Think,* which teaches children counterintuitive concepts in maths & science.
- The programme has shown positive results in a previous study: improvements in science (+2 months of progress), and in maths (+1 month of progress).
- Stop and Think consists of 30 sessions lasting 15 minutes each over 10 weeks and can be easily integrated into regular maths and science teaching.
- This study will run between Autumn 2022 and Summer 2023 and will involve years 3 and 5.
- All schools in the study will receive the software, training for teachers on how to use it and technical assistance for free.

What?

- A software programme developed by neuroscientists at Birkbeck University which uses quizzes and games to help pupils learn counterintuitive concepts in science and maths. For example, making the mistake of thinking that -5 is larger than -1.
- 30 sessions lasting 15 minutes each over the course of 10 weeks (three times a week). Sessions should be integrated into the start of regular maths and science lessons. Participating schools will need a projector/interactive whiteboard.
- In a previous EEF study, pupils who received the programme had higher scores in science (+2 months progress) and in maths (+ 1 months progress).


What number is shown by the blocks?
43
Check Answer
Players
Candice
I have counted seven pieces in total, so that must be the answer.
There are four long bars and three spare cubes.
The four bars each have ten cubes and there are three more units to add.
Candice
Ollie
Kate

Screenshots from the programme.

How?

In this study, we will test the impact of Stop and Think on maths and science attainment in years 3 and 5 using a randomised controlled trial. This means that either year 3 or year 5 at your school will be randomly selected to receive the programme. The other year group will be taught as normal. Randomised controlled trials are considered the strongest type of evaluation to work out if a programme is having an impact.

When?

September 2022 to July 2023. The number of school places is limited so sign up today!

What will it cost my school?

Participation is free. All costs will be covered by the research team.

How much time will be required?

Teachers should use the software in maths and science lessons 3 times a week (15 mins per lesson) over a 10 week period from January 2023 to May 2023. BIT will provide teacher training on how to use the software from October 2022 to January 2023.

How will the programme be evaluated?

Pupils taking part in the study will do a short, age-appropriate assessment (science and maths) at the end of the programme (Summer 2023). External researchers will visit schools to deliver these assessments. We will also gather feedback via a survey and interviews with teachers, and focus groups with students (these will only take place in a small number of schools).

How will data sharing work?

- Participating schools will:
  - Send an information sheet and withdrawal form to parents in September 2022.
  - Provide pupil data (e.g. pupil names, dates of birth, free school meal status), so that we can access their previous results on the National Pupil Database.
- Test results will be linked with information about the pupils from the National Pupil Database (NPD). At the end of the evaluation, the data will be shared with Birkbeck University, the Department for Education, FFT Education (EEF's data processor for their archive), and the Office for National Statistics. Data may also be shared in an anonymised form with other research terms in the future. All EEF trial data is stored in the EEF data archive. The archive does not contain direct identifiers like pupil name, contact details and date of birth, but does hold a Pupil Matching Reference (PMR). The PRM is used for further matching to the NPD and other administrative datasets that may be

required as part of subsequent research. We will not use pupil names or school names in any report arising from this research.

- All pupil information collected as part of the study will be treated with the strictest confidence by the project team in line with the requirements of the GDPR and the Data Protection Act 2018. You can find further information in this NatCen Privacy Notice.

Project team

This is a major study and several different organisations are collaborating on it. They are:

- **The Behavioural Insights Team** (BIT) - *The main organisation your school will have contact with during this study.* BIT is now a social purpose company with offices around the world. BIT applies behavioural science to improve public policy.
- **The National Centre for Social Research** (NatCen) - *Responsible for evaluating the programme and will run the assessments at the end.* NatCen is Britain's largest and oldest social research organisation. They have delivered several school-based studies for charities and the government about what works in education, and are experts at research that involves pupils, young people and teachers.
- **Centre for Educational Neuroscience** (CEN) - *Originally developed the programme.* CEN is a research centre at two world leading universities: Birkbeck and University College London. Academics from CEN originally developed Stop and Think.
- **Education Endowment Foundation** (EEF) - *Funding the study.* The EEF is an independent charity dedicated to breaking the link between family income and educational achievement. They run projects to test the effectiveness of education programmes to improve outcomes for children across the UK.

Key dates/timeline

Month	Activity
Oct 202 - July 2022	Schools sign up to the project (first come, first served!)
Sept 2022	Schools send the information sheet to parents and submit pupil data to NatCen
Oct 2022	NatCen inform schools which year group (year 3 or year 5) will receive the programme
Oct 2022 - Jan 2023	Schools host researcher for short visits to install software and train teachers in how to use it
Jan 2023 - May 2023	Schools use the software in maths and science lessons
	Lesson observations, focus groups and interviews with staff and pupils (only in a small number of schools)
May - July 2023	Teachers complete survey
	Researchers visit school to carry out final assessments
Spring 2024	Study results published

Next steps

We have limited spots for this project. If you are interested in participating or finding out more, please email: stopandthink@bi.team. We look forward to hearing from you soon!

Stop and Think: Learning Counterintuitive Concepts

Information sheet for parents and guardians

Dear parent/guardian,

Over the coming school year, [School Name] is taking part in a research project called *'Stop and Think'*. This letter contains information about the project and what it means for your child. Please read it carefully.

What is Stop and Think?

- Stop and Think is a fun, software-assisted way of helping pupils to learn difficult concepts in maths and science using quizzes and games.
- It will be used at the start of regular maths and science lessons.
- In previous research, Stop and Think was found to increase pupils' attainment in science (+2 months of progress) and maths (+ 1 month of progress).

How the study will work

- It will run between Autumn 2022 and Summer 2023.
- The project involves year 3 and year 5. One of these year groups will be randomly selected to receive the programme while the other will continue as normal.
- To understand if the programme is effective, all pupils in years 3 and 5 will complete short, age-appropriate tests in maths and science in the Summer of 2023.
- Some pupils taking part in the study might also be invited to a focus group or interview to understand what they think of it. You would be contacted separately to provide consent for this closer to the time.

Your child's data

- To run the programme, [School Name] will be required to share information about pupils taking part in the study (e.g. name, date of birth and whether pupils receive free school meals) with NatCen who will conduct analysis to find out if the programme was effective.
- Please read the Privacy Notice for full details.
- We take data security very seriously, and the study will comply fully with the General Data Protection Regulation (GDPR) and the Data Protection Act 2018.
- A report about the study will be published in the Spring of 2024. No individual pupil or school will be identifiable in this report - all details will be fully anonymised.
- If you have any questions about [School Name] sharing data with the organisations running the study, please contact [Relevant person at your school].

How can I withdraw my child from this study?

If you are happy for your child to participate, you do **not need to do anything** but please keep this form for your information.

If you DO NOT want your child's information to be used in the research, please tick the box below, sign and return the attached form to [School Name] by [Date: 2 weeks from date of distribution to parents]. They may still be taught using the Stop and Think software but they will NOT have to take any tests and their information will not be used in the study results.

If you have any questions about the study, please contact:

- StopAndThinkEvaluation@natcen.ac.uk for questions about the evaluation, the way we will use data, or if you decide at any time that you don't want your child's data to be used.
- stopandthink@bi.team for questions about the software or how it will work in your child's school.

_ _

Tick the box below and fill in the details to confirm you DO NOT want to participate in this project.

□

Tick here

Signature: ___________________________________

Name: ___________________________________

Name of your child: ___________________________________

School: [School/College Name]

Date: ___________________________________

Stop and Think Evaluation

Privacy Notice

In line with the UK General Data Protection Regulation (GDPR), we want to inform you how information will be processed in the evaluation of Stop and Think. In this privacy notice, we explain the legal basis for data processing, who will have access to participants' personal data, how data will be used, stored and deleted, and who you can contact with a query or a complaint.

Who's who?

This evaluation is being carried out by independent evaluators, the National Centre for Social Research (NatCen), commissioned by the Education Endowment Foundation (EEF). The Behavioural Insights Team (BIT) will deliver the programme.

You can find out more about NatCen at **www.natcen.ac.uk**.

You can find out more about the EEF at **www.educationendowmentfoundation.org**.

You can find out more about the BIT at **https://www.bi.team/**.

Who will access personal data?

NatCen are carrying out this evaluation and named individuals on the NatCen research team will have access to:

- Sample files provided by schools containing School Unique Reference Numbers (for the National Pupil Database or the NPD) and the following details for all Year 3 and Year 5 pupils:
    1. Unique Pupil Number (UPN) – for the NPD
    2. Date of birth

3. First name
4. Last name
5. Free School Meal (FSM) status
6. Year group
7. Class (if a multi-form-entry school)

- Pupil attainment data from the NPD and pupil assessment data from our impact evaluation (as detailed below)
- School and teachers'/staff members' names and contact details provided by the Behavioural Insights Team
- Audio recordings from pupil focus groups, teacher focus groups and teacher/staff member interviews
- Teacher survey responses

McGowan Transcriptions (**www.mcgowantranscriptions.co.uk**) is the transcription service NatCen use to transcribe our interview and focus group data. They will have access to recordings and transcripts from all interviews and focus groups. McGowan Transcriptions is on NatCen's approved supplier list, and is compliant with all our information security policies.

Formara Print (**www.formara.co.uk**) is the printing company NatCen use. They will print documents containing pupil names, UPNs and dates of birth.

Experienced NatCen interviewers will visit schools to supervise pupil assessments. They will have access to pupil details only for the schools where they supervise assessments.

How will the data be used?

The data collected will be used for research purposes only.

We will a) collect pupil assessment data and b) use attainment data from the National Pupil Database (NPD) in our **impact evaluation**. We will compare results of children who do and do not take part in Stop and Think, to see whether the programme makes a difference to how well they do in maths and science. All assessment and attainment data will be pseudonymised before being analysed.

For the pupil assessments, pupils will complete:

- science and maths progress tests (by GL Assessment) and
- science and maths misconception tests (by NatCen and Oxford MeasurEd).

For the science and maths progress tests, schools will be able to access their school's pupil-level test results via GL Assessment's results portal. For the science and maths misconception tests, NatCen will not share test results with schools.

Information and opinions gathered from pupil focus groups, teacher focus groups, teacher/staff member interviews and teacher surveys will be used in our **process evaluation** to understand how the programme works in practice and what children and schools think of it. All responses will be pseudonymised before being analysed.

All impact evaluation and process evaluation data will be treated with the strictest confidence – no schools, teachers, staff members or pupils will be identified in any report arising from the research.

NatCen will securely delete personal information about participants no more than one year after the evaluation is finished (by September 2025 at the latest).

Assessment results will be linked with information about the pupils from the National Pupil Database (NPD). At the end of the evaluation, the pseudonymised data will be shared with the Department for Education, and with the Office for National Statistics. It will also be stored in the EEF archive (managed by FFT Education) where other research teams can access it. The archive does not contain direct identifiers like pupil names, contact details and date of birth, but does hold a Pupil Matching Reference (PMR). The PMR is used for further matching to the NPD and other administrative datasets that may be required as part of subsequent research.

The legal basis for processing data

For the duration of the evaluation, NatCen is a data controller who also processes data. This means we are responsible for deciding the purpose and legal basis for processing data. The legal basis is "legitimate interest". This means we believe that there is a genuine reason for us to process this data (to evaluate Stop and Think), this data is needed to fulfil this purpose (we could not evaluate Stop and Think without this information) and using this data will not interfere with individuals' interests, rights or freedoms.

After the evaluation ends, data from the impact evaluation will be stored in the EEF archive (as detailed above). At this point, the EEF will become the data controller.

Who can I contact with a query or a complaint?

You have the right to raise any concerns with the Information Commissioner's Office (ICO) via their website at **https://ico.org.uk/concerns**/.

If you are a parent/carer, you have the right to object to your child's information being used in this evaluation. If you do not want your child's information to be used, please contact the Stop and Think Lead at your child's school. You can also email us via **StopAndThinkEvaluation@natcen.ac.uk** or call 0808 164 0397 during office hours.

Contact information

If you have any questions about the evaluation, including how personal information will be processed, please contact the NatCen research team at **StopAndThinkEvaluation@natcen.ac.uk**.

Education
Endowment
Foundation

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn