PROJECT TITLE	**Stop and Think: Learning Counterintuitive Concepts**
DEVELOPER (INSTITUTION)	Birkbeck, University of London; Behavioural Insights Team (BIT)
EVALUATOR (INSTITUTION)	National Centre for Social Research (NatCen)
PRINCIPAL INVESTIGATOR(S)	Mary McKaskill
PROTOCOL AUTHOR(S)	Isabel Taylor, Enes Duysak, Padmini Iyer, Helena Takala
TRIAL DESIGN	Two-arm cluster randomised controlled trial with random allocation at school-level
TRIAL TYPE	Effectiveness
PUPIL AGE RANGE AND KEY STAGE	7 – 10, KS2
NUMBER OF SCHOOLS	165 at analysis (181 at recruitment)
NUMBER OF PUPILS	10,296 at analysis (13,468 at recruitment)
PRIMARY OUTCOME MEASURE AND SOURCE	**Outcome:** Maths attainment amongst FSM pupils **Measure and source:** Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment
SECONDARY OUTCOMES MEASURE AND SOURCE	**Outcomes:** Maths attainment amongst all pupils Science attainment amongst FSM pupils and amongst all pupils Common misconceptions in Maths and Science **Measure and source:** Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment Year 3: Progress Test in Science (PTM8), 0-40, GL Assessment Year 5: Progress Test in Science (PTM10), 0-50, GL Assessment Both years: Age-specific common misconceptions in Maths and Science tests (in development)

Table of Contents

Protocol version history

VERSION	DATE	REASON FOR REVISION
1.0 [*original*]	28/03/2022	-

Study rationale and background

Maths and science skills in the UK

Existing evidence indicates that young people in the UK have relatively poor maths and science skills compared to international standards. Programme for International Student Assessment (PISA) results in 2015 showed that 15-year-olds in the UK ranked 15th in science and 27th in maths, out of 72 participating countries.[1] A 2016 study found that over a quarter of young people aged 16-19 in England had low numeracy skills, with England 22nd out of 23 countries.[2] In a 2014 study, employers felt labour market entrants are not properly prepared for the workforce, with the UK again comparing poorly against other countries in

[1] Organisation for Economic Co-operation and Development (2018) *PISA 2015: PISA results in focus*. Available at: https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf.
[2] 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.

this respect.[3] There is also evidence of intergenerational effects, with poor parental attainment in maths and science reflected in young people's educational outcomes.[4] [5]

The UK government has identified improving Science, Technology, Engineering and Maths (STEM) skills as key to improving the international competitiveness of the UK economy.[6] This is particularly important in light of global technological advances; the automation of increasingly sophisticated tasks through artificial intelligence means that required skills in the labour market are rapidly changing.[7] To meet these needs, successive UK governments have introduced measures to improve attainment in STEM subjects and to increase their take-up at Key Stages (KS) 4 and 5.[8] 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.[9] 'Appropriate, accurate and inspiring' STEM education in primary schools therefore plays a key role in later interest and engagement in STEM subjects and careers.[10]

Learning counterintuitive concepts

Children's ability to learn maths and science concepts may be limited by their ability to inhibit perceptual evidence (what they see, feel or hear) or their pre-existing beliefs.[11] [12] 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.[13] 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.[14] To learn new concepts in maths and science, pupils must be able to inhibit prior contradictory knowledge and misconceptions to acquire new knowledge.[15] Existing evidence suggests this skill varies between pupils, with variation evident from an earlier age and

[3] UK Commission for Employment and Skills (2014) *Employer Perspectives Survey 2014: UK results, Evidence Report 88.* Wath-Upon-Dearne, UK: UK Commission for Employment and Skills.
[4] OECD (2013) *First results from the survey of Adult Skills*. Paris, France: OECD.
[5] 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.
[6] Department for Business, Energy & Industrial Strategy (2017) *Industrial Strategy: building a Britain fit for future.* Available at: https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future
[7] Government Office for Science (2017) *Future of skills and lifelong learning*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/727776/Foresight-future-of-skills-lifelong-learning_V8.pdf.
[8] DfE (2020) *Multi-million government investment in the future of UK science.* Available at: https://www.gov.uk/government/news/multi-million-government-investment-in-the-future-of-uk-science.
[9] Archer Ker, L., Tomei, A. (2013) *What influences participation in science and mathematics?: A briefing paper from the Targeted Initiative on Science and Mathematics Education (TISME).* Available at: https://kclpure.kcl.ac.uk/portal/files/64435093/TISME_briefing_paper_March_2013.pdf.
[10] 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.* Available at: https://www.raeng.org.uk/publications/reports/uk-stem-education-landscape.
[11] Vosniadou et al. (2018) The recruitment of shifting and inhibition in on-line Science and Mathematics tasks. *Cognitive Science* 42(6), 1860-1886. Available at: https://onlinelibrary.wiley.com/doi/full/10.1111/cogs.12624?af=R.
[12] 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.* Available at: doi.org/10.1007/s41465-019-00161-4.
[13] Babai et al. (2015) A warning intervention improves students' ability to overcome intuitive interference. *ZDM Mathematics Education* 47, 735–745. Available at: https://link.springer.com/article/10.1007/s11858-015-0670-y#article-info.
[14] Allen (2014) *Misconceptions in Primary Science.* London: McGraw-Hill Education
[15] NFER (2016) *Protocol for the evaluation of counterintuitive concepts intervention*. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Learning_counterintuitive_concepts_Final.pdf.

weaker control skills among pupils from socio-economically disadvantaged backgrounds compared to their more advantaged peers.[16]

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 was awarded high security (4 padlocks), under EEF's Classification of the Security of Findings.[17]

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 'business as usual' teaching (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.[18]

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.[19] Children who received Stop and Think also made more progress than children in the active control group; 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. 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 outcome of inhibitory control.[20] The post-intervention test used for this intermediate outcome in the efficacy trial required children 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 child from pausing to consider the question before they answer. This focus on speed of response may have impacted these findings.

The efficacy trial showed some initially promising results for children receiving Free School Meals (FSM). On average, FSM pupils made additional progress compared to the control group in (a) Year 3 and Year 5 maths, and (b) Year 5 science.[21] The efficacy trial was not

[16] NFER (2016) *Protocol for the evaluation of counterintuitive concepts intervention*. Available at: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Learning_counterintuitive_concepts_Final.pdf.

[17] EEF (2019) *Classification of the security of findings from EEF evaluations*. Available at: https://educationendowmentfoundation.org.uk/public/files/Evaluation/Carrying_out_a_Peer_Review/Classifying_the_security_of_EEF_findings_2019.pdf.

[18] Inhibitory control: the ability to suppress interfering thought processes or actions. See Carlson, S. M., Moses, L. J., & 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), 73–92. Available at: https://psycnet.apa.org/record/2002-17339-002.

[19] Therefore, in the efficacy trial the evaluators concluded that "the statistical evidence does not meet the threshold set by the evaluator to conclude that the true impact [of the intervention on maths attainment] was nonzero". 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.

[20] 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.

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

powered to measure an effect for FSM pupils, and the evaluators reported that the effects were not statistically significant. However, the post-hoc sensitivity analysis carried out by Durham University suggested a statistically significant effect of the programme on maths attainment amongst FSM pupils that was larger than the estimated effect size on science attainment or maths attainment amongst all pupils.[22] NFER therefore recommended that a subsequent effectiveness trial should further explore impact on children receiving FSM.

Description of the intervention

Name

Stop and Think: Learning Counterintuitive Concepts

Why

Many mistakes in maths and science are made because children tend to answer with an intuitive response based on perceptual evidence or pre-existing beliefs of how the world works. Stop and Think trains children to inhibit their initial response to maths and science problems and give a slower, more reflective answer instead. The rationale is that this will increase children's ability to learn counterintuitive concepts, and thereby increase their attainment in maths and science.

The developers of Stop and Think (Birkbeck, University of London) were motivated by an interest in exploring how new insights from neuroscience research on reasoning, decision-making and inhibition control can be used in education. Emerging research suggests that inhibition control training needs to be embedded in subject-specific knowledge in order to be effective.[23] Birkbeck therefore developed exercises for Stop and Think that are closely related to the maths and science curriculum, rather than general counterintuitive concepts.

Who

The intervention will be delivered by Year 3 and Year 5 teachers in participating primary schools in England. The recipients of the intervention will be the Year 3 and Year 5 pupils. The delivery team at Behavioural Insights Team (BIT) will recruit primary schools to the intervention and support teachers to deliver the intervention. This is a change from the efficacy trial, when the intervention was delivered by the developer team at Birkbeck.

At the recruitment stage, schools will be asked to identify a school lead, who will be responsible for coordinating Stop and Think delivery and evaluation activities. Please see Recruitment (p. 13) and Agreement to participate in the trial (p. 29) for more details.

What

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 programme content includes a series of sessions, made up of questions and multiple-choice answers which include distractors demonstrating common misconceptions. The session topics are aligned to the maths and science curriculum in Years 3 and 5.

Stop and Think has undergone some modifications since the efficacy trial (between September 2020 and March 2022). This has included minor changes to content and software based on: feedback from the efficacy trial; a design workshop with game designers,

[22] 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.
[23] For example: Botvinick, M. M. and Cohen, J. D. (2014) The computational and neural basis of cognitive control: charted territory and new frontier, *Cognitive Science*, 38 (6), pp. 1249–1285. Available at: https://pubmed.ncbi.nlm.nih.gov/25079472/.

teachers and the unLocke team;[24] and a focus group with children to appraise the design of the software. Following modifications based on these consultations, Birkbeck also ran a small-scale validation study in three schools to identify any remaining software issues. See Appendix 1 for full details of changes to Stop and Think since the efficacy trial.

Before the intervention starts, an assistant from BIT will visit schools in October 2022 – January 2023. They will install the Stop and Think software, show the teachers how to use the programme and deliver a short training session on how to deliver Stop and Think sessions. Teachers will need a computer and a projector or an interactive whiteboard to deliver the sessions – this is a requirement for recruitment to the intervention.

Teachers will be provided with a handbook, a briefing video and an FAQ document. Ongoing support will be available from BIT by email and phone if teachers wish to access this. However, unlike the efficacy trial, support will not be proactively offered by the delivery team, based on recommendations from NatCen to replicate more 'real world' conditions during the effectiveness trial. Teachers will still meet an assistant in person during the initial information session, but after this, primary sources of support will be the handbook, briefing video and a frequently asked questions (FAQ) document.

The Stop and Think trial is part of the DfE's 2021-22 Accelerator Fund[25] under the capability-building strand. This additional funding is dedicated to developing an online version of the Stop and Think software as well as an online training model.

How

Stop and Think uses a question and answer format. A character called Andy poses questions to three virtual game-show contestants who demonstrate correct and incorrect thinking based on common misconceptions. Children complete various tasks as if they are taking part in a game show, with the programme actively encouraging them to stop and think before answering. The intervention is delivered to pupils at the start of maths and science lessons by their teacher or teaching assistant.

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

Figure 1: Screenshots from the Stop and Think programme

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

[25] Since the 2021-22 academic year, several EEF projects have benefitted from the DfE's Accelerator Fund. You can read more about the EEF's collaboration with the Accelerator Fund here: https://educationendowmentfoundation.org.uk/support-for-schools/building-the-eefs-pipeline-of-evidence-informed-programmes-through-the-accelerator-fund.

Where

The intervention is delivered in classrooms at participating schools at the start of maths and science lessons.

When and how much

Intervention delivery will start in January 2023 and last until May 2023. This is a change from the efficacy trial, when classroom delivery took place slightly earlier in the academic year (November to March). From a curriculum perspective, this means that pupils will have encountered more of the content covered in the programme by the time delivery starts.

During the delivery period, schools will be expected to deliver a total of 30 Stop and Think sessions, three times per week over a ten-week period. Each session lasts around 12 minutes. The dosage is unchanged from the efficacy trial.

Tailoring

Teachers have some flexibility over how they facilitate the sessions, especially around how the children interact with the software as a group and how they give answers. In addition, there are two optional ways that teachers can tailor their use of the software:

- Teachers can choose weekly themes (e.g. animals, fractions) if they want sessions to fit around their existing curriculum. Alternatively, they can opt for random allocation of themes (which was the default during the efficacy trial).
- Teachers can opt to include motivational elements in the software. One is to include the class in a leader board with other schools, so pupils see how many sessions they have completed, and how they stand in relation to other schools. Pupils will only see the schools immediately above and below them, rather than their overall ranking. Another motivational element is for the class to receive virtual coins for each session completed. These coins can then be used to buy items to improve an animal avatar.

Developers expect these options to encourage use of the software (i.e. increase compliance and dosage), but do not expect teachers' choices about how to use the software themselves to moderate the impacts of the software on children's performance in maths and science. This is because the incentivisation changes take place outside the core Stop and Think sessions and therefore will not influence impacts by affecting the core mechanism of 'stopping and thinking', as set out in the logic model (Figure 2).

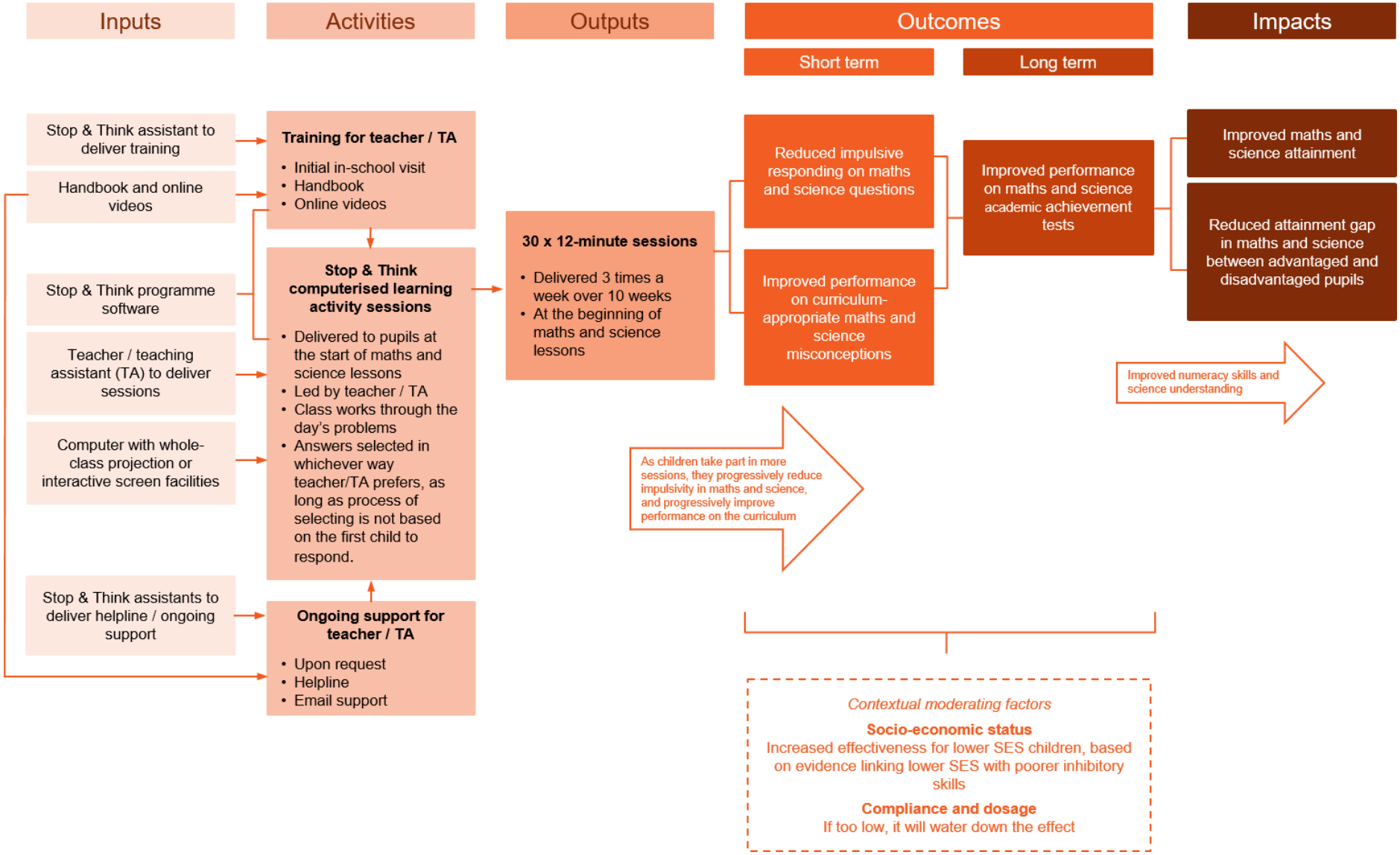
Control condition

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

Logic model

Building on the logic model developed for the efficacy trial, 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



Impact evaluation

Impact evaluation

Research questions

The impact evaluation of Stop and Think will build on findings from the Stop and Think efficacy trial and subsequent sensitivity analysis that found that the Stop and Think programme had a significant effect on maths attainment amongst FSM pupils, and that this effect was larger than for any other outcome tested. As the efficacy trial also found the Stop and Think programme to have a significant effect on pupil's science attainment, this (along with maths attainment amongst all pupil) will also be included as a secondary outcome in this trial.

This impact evaluation therefore aims to answer the following research questions:

Primary research question

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

Secondary research questions

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

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

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

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

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

Design

The evaluation will be 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. A randomised controlled trial (RCT) uses the mechanism of randomisation to assess the causal impact of an intervention. Randomisation, if conducted correctly, should result in there being no important differences at baseline between treatment and control groups in the main determinants of our outcomes of interest. Any differences at baseline are due to chance and are accounted for in the statistical analysis. As a result, any discrepancy in outcomes at the end of the trial can be attributed to the intervention itself. As an effectiveness trial, the evaluation aims to test the effect of the intervention in 'real-world' circumstances.

Stop and Think is designed to improve Year 3 and Year 5 maths and science attainment by improving pupils' ability to adapt to counterintuitive concepts. The primary outcome of interest is maths attainment among Year 3 and Year 5 pupils from disadvantaged backgrounds (as defined as those who have been eligible for FSM at any point in the previous 6 years). This is the primary outcome for this trial as it is a key priority for EEF to improve attainment in maths among pupils from disadvantaged backgrounds. Furthermore, although the logic model does not imply that the intervention will have a larger impact on maths or science attainment and the efficacy trial found similar sized effects for both maths

and science attainment, post-hoc analysis of the efficacy trial found that Stop and Think had the largest impact on maths attainment amongst FSM-eligible pupils in models estimated for pupils in Year 3 and Year 5 combined.[26]

The secondary outcomes include maths attainment for all Year 3 and Year 5 pupils, science attainment for all Year 3 and Year 5 pupils, science attainment for Year 3 and Year 5 pupils from disadvantaged backgrounds and the prevalence of common misconceptions in maths and science among all Year 3 and Year 5 pupils. We will measure maths and science attainment following intervention delivery by administering age-specific GL Progress Tests in Maths and Science (PTM8 and PTS8 for pupils in Year 3 and PTM10 and PTS10 for pupils in Year 5). We will measure common misconceptions using age- and subject-specific tests being developed by NatCen. Individual pupils will be randomly assigned to sit either maths or science tests, so that 50% of pupils in each year group will be tested in maths attainment and maths misconceptions and 50% in science attainment and science misconceptions. More details on the outcome measures are provided on p.17.

Due to the disruption in national curriculum testing caused by the COVID-19 pandemic, identical baseline measures for both Year 3 and Year 5 pupils involved in the trial will not be available. The trial will therefore use KS1 maths scores as a measure of prior attainment for pupils in Year 3 and the Early Years Foundation Stage Profile (EYFSP) point score as a measure of prior attainment for pupils in Year 5. More details on these baseline measures are provided on p. 17.

The trial has been designed to detect an effect of 0.17 on maths attainment among KS2 pupils from disadvantaged backgrounds, as the primary outcome. This will ensure that the trial is sufficiently powered to detect an effect of the size found in the post-hoc analysis of the effect of Stop and Think on maths attainment amongst FSM pupils in the efficacy trial.[27] This requires a final analytical sample of 165 schools (see Appendix 2). To allow for the anticipated drop out of schools after recruitment (including the period up to and after randomisation) we aim to recruit at least 181 schools.[28] Year 3 and Year 5 pupils in these schools will be the trial participants.

We will randomly assign schools into two conditions. In each condition, only one year group will receive the intervention while the other year group will follow the 'business-as-usual' teaching. Such that, for schools randomly assigned into condition 1, Year 3 pupils will receive the intervention of the Stop and Think computer programme in the academic year 2022-23, while Year 5 pupils will receive 'business as usual' classes. For each school assigned to condition 2, Year 3 pupils will receive 'business as usual' classes, while Year 5 pupils will receive the Stop and Think programme.

[26] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts. Unpublished manuscript.
[27] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts. Unpublished manuscript.
[28] We aim to recruit at least 180 schools, with an expectation that around 9% of schools will likely dropout between recruitment and analysis (either before or after randomisation), leaving around 164 schools at the post-intervention analysis stage. This is based on the efficacy trial of the Stop and Think, where 97 schools were recruited and 11% of these schools withdrew from the programme for various reasons, leaving 87 primary schools for the analysis. See 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.

Table 1: Trial design

Trial design, including number of arms		Two-arm, cluster randomised control trial
Unit of randomisation		School level
Stratification variables (if applicable)		School-form entry and the school-level proportion of pupils eligible for FSM
Primary outcome	variable	Maths attainment amongst FSM pupils
	measure (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment; Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment
Secondary outcome(s)	variable(s)	Maths attainment amongst all pupils Science attainment amongst FSM pupils and amongst all pupils Common misconceptions in maths and science amongst FSM pupils and amongst all pupils
	measure(s) (instrument, scale, source)	Year 3: Progress Test in Maths (PTM8), 0-55, GL Assessment; Year 5: Progress Test in Maths (PTM10), 0-65, GL Assessment Year 3: Progress Test in Science (PTS8), 0-40, GL Assessment; Year 5: Progress Test in Science (PTS10), 0-60, GL Assessment Both years: Age-specific common misconceptions in maths and science tests
Baseline for primary outcome	**variable**	Year 3: Maths attainment Year 5: Overall attainment
	measure (instrument, scale, source)	Year 3: KS1 maths attainment, 8-category variable ranging from BLW (below expected standard) to GDS (working at a greater depth), National Pupil Database Year 5: Overall EYFSP Point Score, 1-3, National Pupil Database
Baseline for secondary outcome	**variable**	Year 3: Maths attainment Year 5: Overall attainment
	measure (instrument, scale, source)	Year 3: KS1 maths attainment, 8-category variable ranging from BLW (below expected standard) to GDS (working at a greater depth), National Pupil Database Year 5: Overall EYFSP Point Score, 1-3, National Pupil Database

Randomisation

Every school involved in the trial will have one intervention and one control year group, ensuring equal allocation of condition 1 and condition 2. School-level randomisation will assign Year 3 in each school to treatment or control, with Year 5 in the same school assigned to the opposite condition. So, for example, in a school in which Year 3 is randomly assigned to treatment, Year 5 will be allocated to control. This means that 50% of schools will have Year 3 allocated to treatment and Year 5 allocated to control, while the remaining

50% of schools will have Year 5 allocated to treatment and Year 3 allocated to control. Before randomisation, schools will be stratified by class-form entry size (whether there is 1, 2 or 3+ classes per year group per year) and the school-level proportion of FSM-eligible pupils[29] prior to randomisation to ensure balance across treatment and control groups.

This approach has two main benefits. Firstly, it provides greater power to detect an effect compared to overall school-level randomisation for both year groups combined. Secondly, it helps to maintain school engagement. No schools will be assigned solely to the control condition, and so every school will receive the intervention.

We do not anticipate serious contamination issues between year groups within the same school with this randomisation approach, as teachers in primary schools tend to be allocated a single class. The efficacy trial utilised the same randomisation approach for 87 primary schools in England and no contamination issues were identified. We will monitor the control condition via the implementation and process evaluation.

Randomisation will be carried out blind by an analyst at NatCen in autumn term 2022. Randomisation will be undertaken in Stata and both do and log files will be used to record the randomisation process. At time of randomisation, analysts will be blinded to school identity. School identifiers will then be merged with group allocation data after randomisation.

For testing purposes, pupils will be randomised to take either the maths or science tests. This approach will ensure that pupils and schools will not be over-burdened by requiring all pupils to take two sets of attainment and misconceptions tests (in both maths and science) while individual-level randomisation provides greater power than school randomisation to testing in either subject. This means that 50% of pupils will be randomly allocated to take the age-specific GL Progress Test in Maths and the maths misconceptions test and 50% to take the age-specific GL Progress Test in Science and the science misconceptions test.

Recruitment

BIT will identify and recruit eligible schools for this trial. All state primary schools in England will be eligible for the trial unless a) pupils from Year 3 and Year 5 are taught in the same class; b) the school has previously partnered with the Stop and Think programme; or c) the school has taken part in piloting work for the misconceptions tests which are being used as a secondary outcome measure.

Recruitment of schools will take place between November 2021 and July 2022. Schools will need to sign a Memorandum of Understanding (MoU) confirming their commitment to delivering the programme as required and taking part in evaluation activities. This process is expected to be completed before the start of the 2022-23 academic year. Information sheets shared with schools and parents are available in appendices 5 and 6.

As we expect schools to drop-out between recruitment and analysis (both before and after randomisation) at a similar level to that measured in the efficacy trial,[30] we aim to recruit a higher number of schools to ensure that the analytical sample remaining for post-intervention testing can power this study to a suitable MDES. The recruitment target for schools therefore assumes that 9% of the recruited sample of eligible schools will drop out ahead of analysis.

[29] Three categories of schools will be created based on their school-level proportion of FSM-eligible pupils, with boundaries set at the terciles of the distribution to create comparatively sized groups reflecting a low, medium and high proportion of FSM-eligible pupils.

[30] In the efficacy trial, 97 schools were originally recruited, though five of these were not eligible for the trial as they did not teach Year 3 and Year 5 pupils in different classes. The efficacy trial therefore recruited 92 eligible schools with 87 schools remaining in the trial at randomisation and three schools lost prior to post-intervention testing. 84 schools were therefore included the final sample for analysis, meaning that 8 eligible schools (or 9%) were lost after recruitment. 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.

Participants

All Year 3 and Year 5 pupils in recruited schools will be eligible for the trial. Participating schools will be asked to share the trial information leaflet and privacy notice with all Year 3 and Year 5 pupils and their parents/carers in September 2022. The parents/carers will be given two weeks to withdraw from the trial (and from data processing). After this period, schools will be asked to share pupil-level information for all pupils in Year 3 and Year 5 who have not been formally withdrawn from the trial with NatCen. These pupils will make up the trial participants.

Schools will be required to provide background information for all the trial participants in October 2022. This will include the 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 will be collected in an Excel spreadsheet template and uploaded by schools using a secure NatCen website.

From the sample of trial participants, we expect an average year group size of 37.2 pupils by primary school year. With an expected pupil attrition of 16%,[31] we assume an average of 31.3 per year group per school to be included in the final sample for analysis. As in the previous efficacy trial, pupils who were eligible for FSM at any point over the previous 6 years will be used to identify FSM eligibility. With a current national average of 22.1% of pupils having been eligible for FSM at any point over the previous 6 years,[32] we estimate an average of 6.9 pupils per year group per school in the final sample for analysis to be from disadvantaged backgrounds.

As explained in the Randomisation section on p.12, half of the pupils will be randomly allocated to take maths tests and half to take science tests. We therefore assume an average of 15.6 pupils per year group per school (of which 3.5 are estimated to be FSM eligible) will be tested in maths and science respectively.

Power calculations

We have designed this trial to be powered to detect an effect of the size found in the Durham Univeristy's post-hoc analysis of the efficacy trial[33] 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). Full details of the power calculations determining the minimum sample sizes required at analysis are included in Appendix 2. Power calculations for maths and science attainment for the estimated sample at recruitment are included in Appendix 3.

Power calculations have been estimated for an approach used in individual participant data (IPD) meta-analysis, as used by Ashraf et al (2021) in their IPD analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals.[34] This approach is considered to be the "gold standard for meta-analysis".[35] It will allow us to estimate a single effect size for pupils across both Year 3 and Year 5, by treating each year group (each of which has different baseline and outcome measures) as, in effect, participants in separate trials.

[31] Based on the pupil retention rate from the efficacy trial of the 'Stop and Think' programme. Roy, P. et al. (2019) *Stop and Think: Learning counterintuitive concepts. Evaluation report,* p 33. Available at: https://www.nfer.ac.uk/media/3703/learning_counterintuitive_concepts_evaluation_report_-final.pdf.
[32] Calculated from DfE (2022) *Schools, Pupils and their Characteristics: January 2021.* London: DfE
[33] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts. Unpublished manuscript.
[34] Ashraf et al. (2021). *Individual participant data meta-analysis of the impact of EEF trials on the educational attainment of pupils on Free School Meals: 2011 – 2019.* EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/documents/evaluation/evaluation-syntheses/Individual-participant-data-meta-analysis-of-the-impact-of-EEF-trials-on-the-educational-attainment-of-pupils-on-Free-School-Meals.pdf
[35] MRC. (2020). Individual Participant Data Meta-analyses. Available at: https://www.ctu.mrc.ac.uk/aboutus/our-structure/individual-participant-d

For this approach, power calculations have been estimated for a three-level model. This takes into account that, under this approach, pupils are nested in schools, in years (or trials). It does not include an additional level for classes for two reasons. First, just over half of primary schools in England have only one class per academic year group; the average number of Key Stage 2 classes in state-funded primary schools in England is 5.7, with 51% having only one class per year.[36] Second, only a small minority of primary schools teach classes in ability groups (12% across all subjects, and 19% for maths alone), meaning that class-level clustering in primary schools tends to be much weaker than that measured in secondary schools (Demack, 2019)[37].

In addition to this, the approach used in the efficacy trial will also be used as a sensitivity analysis. For this, separate models will be estimated for each year group, taking into account that pupils (in each year group) are nested within schools. The effect sizes of these models will be combined, using the formula in Borenstein, Hedges, Higgins, & Rothstein, (2009, p. 66) for independent samples. This will allow us to estimate the overall effect of the Stop and Think programme on maths attainment and on science attainment for Year 3 and Year 5 pupils combined using a comparable method to that used in the efficacy trial.

We estimated the required sample sizes for an IPD meta-analysis approach for a Minimum Detectable Effect Size equal to the effect sizes detected in the efficacy trial and subsequent post-hoc analysis. We assume that the intra-cluster correlations and the pre-test and post-test correlations between EYFSP scores and maths attainment will be in line with those measured in the efficacy trial[38], and use published figures from the FFT Education Datalab to estimate post-test correlation between KS1 attainment and KS2 attainment[39]. Data from the DfE's statistics on current pupil numbers is used to estimate the number of pupils in the relevant year groups.[40]

Table 2 below shows the power calculations for maths attainment (the primary outcome) for the analytic sample estimated for this study. Table A2 in Appendix 2 shows the power calculations used to estimate the minimum required sample sizes to power the trial to detect the effects found in the efficacy trial on maths attainment and on science attainment for all pupils and FSM pupils (covering both primary and secondary outcomes). This shows that 165 schools would be required to power the trial for maths attainment among FSM pupil. Therefore, a final analytical sample of 165 schools at analysis will be sufficient to power the trial for maths attainment among FSM pupils (our primary outcome and population of interest) and maths attainment among all pupils. It would also power the trial for an MDES of 0.14 for science attainment among all pupils, smaller than the effect size detected in Durham University's post-hoc analysis[41], as shown in Table A3 in Appendix 2.

To allow for school losses after recruitment, as explained in the Recruitment section on p. 13, we aim to recruit 181 schools assuming that 9% will be lost from the trial after recruitment, meaning 165 schools will be included in the post-intervention analysis. The sample size calculations included in Table A4 in Appendix 3 reflect a larger number of schools and pupils as they are based on the number of schools and pupils estimated at recruitment.

[36] Calculated from DfE (2022) *Schools, Pupils and their Characteristics: January 2021.* London: DfE
[37] Demack, S. (2019). Does the classroom level matter in the design of educational trials? A theoretical & empirical review. EEF. Available at: http://shura.shu.ac.uk/25753/1/Does_the_classroom_level_matter.pdf
[38] Roy, P. et al. (2019) *Stop and Think: Learning counterintuitive concepts. Evaluation report, p 34.* Available at: https://www.nfer.ac.uk/media/3703/learning_counterintuitive_concepts_evaluation_report_-final.pdf.
[39] From FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One. London: FFT E.D, p 34
[40] Calculated from DfE (2022) *Schools, Pupils and their Characteristics: January 2021.* London: DfE
[41] The number of schools that would be required to power the trial for science attainment among FSM pupils is 321 schools, which is considered to be too large to be feasible for recruitment. More details can be found in Appendix 2.

As outlined in the Participants section on p. 14, we assume an average of 37.2 pupils per year group per school will be recruited to this trial. With an expected pupil attrition of 16%,[42] we assume an average of 31.3 pupils per year group per school to be included in the final sample for analysis. As outlined in the Randomisation section on p. 12, 50% of pupils (an average of 18.6 recruited pupils per year group per school and an average of 15.6 pupils per year group per school for analysis) will be tested each in maths and in science after the intervention is completed.

As 22.1% of pupils in KS2 in state primary schools in England are currently eligible for FSM, we estimate an average of 8.2 FSM eligible pupils will be recruited to the trial per year group per school, with an average of 4.1 FSM eligible pupils (as recruited) each tested in maths and in science per year group per school. Given the pupil attrition rate explained above, we estimate an average of 6.9 FSM eligible pupils per year group per school at analysis thus an average of 3.5 FSM eligible pupils per year group per school to be tested in each maths and science.

Table 2: Power calculations for maths attainment for estimated analytical sample

		All pupils	FSM
Minimum Detectable Effect Size (MDES)		0.13	0.17
Pre-test/ post-test correlations	level 1 (pupil)	0.635	0.635
Intracluster correlations (ICCs)	level 2 (school)	0.07	0.07
Alpha		0.05	0.05
Power		0.8	0.8
One-sided or two-sided?		2	2
Average cluster size		15.6	3.5
Number of years per school		2	2
Number of year groups	Treatment[43]	165	165
	Control	165	165
	Total	330	330
Number of schools	Treatment[44]	165	165
	Control	165	165
	Total	165	165
Number of pupils	Treatment	2574	578
	Control	2574	578

[42] Based on the pupil retention rate from the efficacy trial of the 'Stop and Think' programme. Roy, P. et al. (2019) *Stop and Think: Learning counterintuitive concepts. Evaluation report,* p 33. Available at: https://www.nfer.ac.uk/media/3703/learning_counterintuitive_concepts_evaluation_report_-final.pdf.

[43] We will randomly assign schools to two conditions. In each condition, one year group will receive the intervention while the other year group will follow the 'business-as-usual' teaching. Condition 1 will involve Year 3 pupils receiving the intervention while Year 5 pupils receiving 'business as usual' teaching. Schools assigned to condition 2 will involve Year 3 pupils receiving 'business as usual' classes, while Year 5 pupils receiving the intervention. Therefore, every school involved in the trail will have one intervention and one control year-group. Treatment group in this table reflects the condition 1 while control group reflects the condition 2.

[44] Due to the randomisation model outlined above, every participating school will have both treatment and control groups.

	All pupils	FSM
Total	5148	1156

Power calculations for the primary outcome are estimated using the following assumptions: KS1 maths attainment (for Year 3 Pupils) and EYFSP overall point score (for Year 5 pupils) are used as individual-level covariates. The age-specific GL Assessment Progress Test in Maths is used as the primary outcome measure for both year groups. The correlation between KS1 maths attainment and GL Progress Test in Maths result is estimated to be 0.76[45] while the correlation between EYFSP overall point score and this Progress Test in Maths result is estimated to be 0.51.[46] The school-level intra-cluster correlation is assumed to be 0.07 for this primary outcome.[47]

The calculations were undertaken using 'PowerUp!'[48] and indicate that the trial is powered to detect an effect of 0.17 standard deviations for the primary analysis for maths attainment among KS2 pupils from disadvantaged backgrounds.

Baseline measures

KS1 maths attainment will be used as the baseline measure of attainment for Year 3 pupils, assuming KS1 tests go ahead as planned in 2021/22.[49] As national testing in schools was cancelled due to the Covid-19 pandemic in 2019/20, when these children were 5 years old, no EYFSP data is available for this cohort. Therefore, if KS1 testing is disrupted in 2021/22, additional pre-testing using age-appropriate GL Assessment tests will need to be considered.

The cancellation of national testing in 2019/20 also prevents us from using KS1 test scores as prior attainment measures for the Year 5 cohort in this trial. We will therefore use the EYFSP overall point score[50] as the baseline measure of attainment for Year 5 pupils.

We will obtain these baseline measures and FSM eligibility information for all trial participants from the National Pupil Database (NPD) after we receive the pupil sample in autumn term 2022.

Outcome measures

As outlined above, 50% of pupils in the trial will be randomly allocated to take maths tests and 50% to take science tests. This will ensure that pupils and schools are not over-burdened with testing in both subjects, while offering greater power than school-randomisation for testing.

[45] Based on correlation between KS1 and KS2 attainment from FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One London: FFT E.D.
[46] Based on correlation between EYSFP and KS2 maths attainment found in the efficacy trial. 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.
[47] As found in the efficacy trial. 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.
[48] Maynard, R. A. & Dong, N. (2013) PowerUP!: A Tool for Calculating Minimum Detectable Effect Sizes and Minimum Required Sample Sizes for Experimental and Quasi-Experimental Design Studies. Available from: https://repository.upenn.edu/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1265&context=gse_pubs.
[49] We will use a variable named "KS1_MATH_OUTCOME" from the National Pupil Database (NPD) as the baseline measure for Year 3 pupils. This is a categorical variable consisting of 8 categories reflecting the level of mathematical knowledge. The categories are as follows: BLW (below expected standard), PKF (Pre-Key stage – Foundations for the expected standard), PK1 (Pre-Key stage standard 1), PK2 (Pre-Key stage standard 2), PK3 (Pre-Key stage standard 3), PK4 (Pre-Key stage standard 4 (Working towards the expected standard)), EXS (working at the expected standard), and GDS (working at a greater depth within the expected standard).
[50] Following the efficacy trial, we will combine all 17 early learning goals to obtain an average EYFSP point score. These variables will be obtained from the National Pupil Database (NPD) and take a value between 1 and 3, where higher scores reflect higher attainment for the specific learning goal,

As there is no theoretical reason that the Stop and Think intervention should have a bigger impact on either maths or science attainment, both were used as co-primary outcomes in the efficacy trial. However, for effectiveness trials, EEF insists on a single primary outcome. Analysis in the efficacy trial found a positive but non-significant effect of the Stop and Think programme on FSM pupil's Maths attainment in separate models for year 3 and year 5 pupils. Furthermore, although this was not included in the published analysis of the efficacy trial, routine post-hoc analysis carried out by the Durham University found a significant, and comparatively large, impact of the intervention for Year 3 and Year 5 FSM-eligible pupils[51]. In addition, addressing pupil disadvantage is a key priority area for EEF. Therefore, maths amongst FSM-eligible pupils was selected as the single primary outcome for this effectiveness trial. The **primary outcome measure** is a standardised measure of pupils' mathematical skills and knowledge, the GL Progress Test in Maths (GL PTM)[52] for pupils from disadvantaged backgrounds.

Age-appropriate versions of the paper-based test will be delivered to the two year groups (PTM8 for Year 3 and PTM10 for Year 5). Pupils will be assessed in May-July 2023.[53] These tests will be used as no relevant national tests are available for Year 3 and Year 5 pupils through the National Pupil Database, so GL Progress Test are appropriate age-specific tests for Maths and Science outcomes.

NatCen will use GL Progress Test in Science (GL PTS), a standardised measure of pupils' science skills and knowledge as a **secondary outcome measure**. Age-appropriate versions of this test will be delivered to the two year groups (PTS8 for Year 3 and PTS10 for Year 5). Pupils' science and maths attainment will be assessed at the same time, in May-July 2023.

The efficacy trial of the Stop and Think programme also explored the effect of the intervention on a general measure of inhibitory control as a secondary outcome. However, the trial did not find an impact on this outcome measure. As the post-intervention test used for this outcome in the efficacy trial asked children to respond to as many questions as possible within a set time-limit, this focus on the speed of response may have impacted the findings. Therefore, for the effectiveness trial, in line with the Stop and Think logic model, NatCen will develop new tests for common misconceptions in maths and science used as outcome measures in additional models to estimate the effect of Stop and Think on intermediate outcomes. Four different tests will be developed for this analysis (one test per subject per year group). Please see Appendix 4 for more details on test development.

During endline testing, pupils will take age-appropriate tests for common misconceptions in maths or science, depending on their randomised allocation to either maths or science attainment tests as outlined above. The 50% of pupils randomised to take the GL Progress Test in Maths will also take the common misconceptions test in maths. The 50% randomised to take the GL Progress Test in Science will take the common misconceptions test in science. A technical note on the misconception tests will be added as an appendix to this protocol once the tests have been developed.

Analysis

A full statistical analysis plan (SAP) will be written and within three months of the completion of the randomisation process place and before any outcome data is seen. This will be published in addition to this protocol.

Primary analysis

The primary analysis will estimate the intervention's impact on maths attainment, as measured by the GL Progress Test in Maths (GL PTM), among pupils from disadvantaged

[51] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts (137). Unpublished.
[52] For more information, please see https://www.gl-assessment.co.uk/products/progress-test-in-maths-ptm/.

backgrounds. In line with the EEF analysis guidance,[54] the primary analysis will follow an intention-to-treat (ITT) approach as this will reflect the real-world conditions in which the programme would be delivered. The analysis of the impact of the intervention will be conducted at the pupil level, with the average maths score from the intervention group compared with the average maths scores from the control group.

As the primary focus of this project is to measure impact of the intervention on maths attainment of pupils coming from disadvantaged backgrounds, the primary analysis will be a subgroup analysis including only these pupils. As suggested by the EEF analysis guidance,[55] the variable "EVERFSM_6_P_[term][yy]" from the NPD, which indicates if a pupil has been recorded as eligible for FSM at any time in the last 6 years, will be used to identify pupils coming from a disadvantaged background.

An approach used in IPD meta-analysis will be followed to build a combined model for both year 3 and year 5 pupils. This will include an additional (third) level, to reflect pupils, nested within schools, nested within year groups (or trials). From this model, we will estimate a single overall effect for both year 3 and year 5 pupils combined for the primary outcome. The standardized PTM score from age-appropriate tests will be used as the dependent variable for both years in this model. This model will include standardized pre-trial test scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate.[56]

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

$$PTM_{ijk} = \beta_{0k} + \beta_{1k}Baseline_{ijk} + \beta_2 Intervention_{jk} * \beta_3 Year\ group_k + u_{jk} + v_k + e_{ijk}$$

where pupils eligible for FSM (i) are clustered within schools (j) within trials (k). The intervention effect is estimated by β_2, and the moderating effect of year group is estimated by $\beta_2 * \beta_3$. The term u_{jk} is a year group-level random effect, v_k is a school-level random effect, and e_{ijk} is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. The stratification variables will be included as fixed effects in this model, since the school-level random effects will control for both observed and unobserved school-level characteristics, including stratification variables.[57] In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 16 SE-64.

The impact of the intervention will be expressed as a standardised effect size using Hedge's g with 95% confidence intervals. Following EEF guidelines, the unconditional variance in the primary outcome for the pooled sample will be used when computing the Hedge's g statistic.

The approach used in the efficacy trial will also be used as a sensitivity analysis. For this, a two-level fixed effects model will be built for each year group with the raw PTM score as the dependent variable. The model will reflect the structure of the data with pupils nested within schools. Each model will include pre-trial test score (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate.[58]

The basic form of the model for pupils eligible for FSM in each year group is:

$$PTM_{ik} = \beta_0 + \beta_1 Baseline_{ik} + \beta_2 Intervention_k + v_k + e_{ik}$$

[54] EEF (2018) Statistical analysis guidance for EEF evaluations.
[55] EEF (2018) Statistical analysis guidance for EEF evaluations.
[56] We will use a variable named "KS1_MATH_OUTCOME" from the NPD as the baseline measure for Year 3 pupils. Following the efficacy trial, we will use average EYFSP point score, which will be formed by combining all 17 early learning goals, as the baseline measure for Year 5 pupils.
[57] Schools will be stratified by class-form entry size and the school-level proportion of pupils eligible for FSM at any time during the past 6 academic years.
[58] We will use a variable named "KS1_MATH_OUTCOME" from the NPD as the baseline measure for Year 3 pupils. Following the efficacy trial, we will use average EYFSP point score, which will be formed by combining all 17 early learning goals, as the baseline measure for Year 5 pupils.

where pupils eligible for FSM (i) are clustered within schools (k). The intervention effect is estimated by β_2. The term v_k is a school-level fixed effect, and e_{ik} the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. The stratification variables will not be included as a covariate in this model, since the school-level fixed effects will control for both observed and unobserved school-level characteristics, including stratification variables.[59] In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 16 SE-64.

To estimate a single effect size for both Year 3 and Year 5 pupils from these separate models, the mean of the two resulting effect sizes will be taken to calculate a single effect size. The effect size calculated in this way will be comparable with findings from other studies, including the Stop and Think efficacy trial. The variance of the combined effect size will be estimated using the formula in Borenstein, Hedges, Higgins, & Rothstein, (2009, p. 66)[60].

As further sensitivity analysis, an alternative model will be estimated to assess whether the findings for the primary analysis are robust to different model specifications. We will estimate the impact of the programme on the sample as a whole with a three-level model including an interaction term between the treatment status and a dummy variable indicating FSM status. Furthermore, if differential loss to follow-up creates an imbalance between trial groups or if attrition is high, the sensitivity of the estimated effect will be assessed by approximating missing outcomes using multiple imputation.

Compliance analysis

The Complier Average Causal Effect (CACE)[61] will be estimated to show the impact of Stop and Think on the primary outcome (maths attainment among KS2 pupils from disadvantaged backgrounds) compared to individuals in the control group, taking into account level of compliance with Stop and Think.

Data for our compliance analyses will be collected during the implementation through the computer programme that delivers Stop and Think. This will be used to measure if and how fully the intervention has been delivered to classes.[62] Specifically, the computer programme will count the number of sessions delivered to each class (from 0 to 30). Following the approach used in the efficacy trial, we will use the number of sessions delivered to each class as a continuous measure of compliance. We will assume one-sided non-compliance in our analysis as we assume that none of the pupils in the control group can be exposed to these sessions since the control group receive a business-as-usual teaching approach and do not have access to Stop and Think.

We will estimate the CACE using a two-stage least square (2SLS) method (Angrist and Imbens, 1995) with the treatment allocation as the instrumental variable for the compliance measure.

Secondary analyses

We will estimate the impact of the programme on attainment in science (amongst all pupils and amongst FSM pupils) and maths (amongst all pupils) using the same IPD meta-analysis approach as outlined for the primary outcome above. For each secondary outcome, three-level models will be estimated for both year 3 and year 5 pupils combined, to reflect pupils, nested within schools, nested within year groups (or trials). The standardized PTM or PTS

[59] Schools will be stratified by class-form entry size and the school-level proportion of pupils eligible for FSM at any time during the past 6 academic years.
[60] Borenstein, M., Hedges, L. V., Higgins, J. P T., & Rothstein, H. R. (2009). Introduction to meta-analysis. London: Wiley, pp. 66
[61] Corresponding to the average effect of the intervention for those pupils who have complied with the programme.
[62] Additional data will be available from the developer on the average amount of time per session, the spacing of sessions and the number of structured practice activities completed.

score will be used as the dependent variable in these models. This model will include standardized pre-test scores (KS1 maths outcome for Year 3 pupils and the EYFSP overall points score for Year 5 pupils) as a covariate.[63]

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

$$Outcome_{ijk} = \beta_{0k} + \beta_{1k}Baseline_{ijk} + \beta_2 Intervention_{jk} * \beta_3 Year\ group_k + u_{jk} + v_k + e_{ijk}$$

where pupils (i) are clustered within schools (j) within trials (k). The intervention effect is estimated by β_2, and the moderating effect of year group is estimated by $\beta_2 * \beta_3$. The term u_{jk} is a year group-level random effect, v_k is a school-level random effect, and e_{ijk} is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. The stratification variables will be included as fixed effects in this model, since the school-level random effects will control for both observed and unobserved school-level characteristics, including stratification variables.[64] In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 16 SE-64

Sensitivity analyses

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

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

$$Outcome_{ik} = \beta_0 + \beta_1 Baseline_{ij} + \beta_2 Intervention_j + v_j + e_{ij}$$

where pupils (i) are clustered within schools (j). The intervention effect is estimated by β_2. The term v_j is a school-level fixed effect, and e_{ij} the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. For these measures we will also be using standardised scores (z-scores), reporting confidence intervals at 95% level, and the effect size utilizing hedge's formula as previously described.

To estimate a single effect size for both Year 3 and Year 5 pupils for each outcome from separate models for each year group, the mean of the two resulting effect sizes will be taken to calculate a single effect size that is comparable with findings from other studies, including the efficacy trial. The variance of the combined effect size will be estimated using the formula in Borenstein, Hedges, Higgins, & Rothstein, (2009, p. 66)[65].

To measure the effects of Stop and Think on misconceptions in maths and science, we will estimate the effects in a similar way to how we estimate the impact of the programme on the primary outcome. The model will follow an intention-to-treat approach but will not include any baseline measures of misconceptions in maths and science. As we do not propose sub-group analysis by FSM status for misconceptions in maths and science, the analysis will include the FSM status as a covariate. The basic form of the model for both year groups combined is as follows:

[63] We will use a variable named "KS1_MATH_OUTCOME" from the NPD as the baseline measure for Year 3 pupils. Following the efficacy trial, we will use average EYFSP point score, which will be formed by combining all 17 early learning goals, as the baseline measure for Year 5 pupils.
[64] Schools will be stratified by class-form entry size and the school-level proportion of pupils eligible for FSM at any time during the past 6 academic years.
[65] Borenstein, M., Hedges, L. V., Higgins, J. P T., & Rothstein, H. R. (2009). Introduction to meta-analysis. London: Wiley, pp. 21-32

$$Outcomes\_Misconception_{ijk} = \beta_{0k} + \beta_1 Intervention_{jk} * \beta_2 Year\ group_k + +\beta_3 Pupil\_FSM\_Status_{ijk} + u_{jk} + v_k + e_{ijk}$$

where pupils are clustered within schools (j) within trials (k). The intervention effect is estimated by β_1, and the moderating effect of year group is estimated by $\beta_1 * \beta_2$. $Pupil\_FSM\_Status_k$ represents whether a pupil has ever been eligible for FSM at any time during the past 6 academic years. The term u_{jk} is a year group-level random effect, v_k is a school-level random effect, and e_{ijk} is the error term, assumed to be normally distributed and uncorrelated with all the covariates included in the model. The stratification variables will be included as fixed effects in this model, since the school-level random effects will control for both observed and unobserved school-level characteristics, including stratification variables.[66] In line with the EEF analysis guidance, other additional covariates will not be considered at this stage. The analysis will be implemented in Stata 16 SE-64.

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

Implementation and process evaluation

Research questions

An implementation and process evaluation (IPE) will be carried out to address the research questions and evaluation domains set out in Table 3. Our IPE domains of interest are informed by EEF's framework for implementation and process evaluation.[67]

Table 3: IPE research questions and domains

Research question	Domain
1. To what extent do the delivery partners and teachers deliver Stop-and-Think as intended?	*Fidelity*
2. How well is Stop-and-Think delivered?	*Quality*
3. How, why and to what extent are changes made to Stop-and-Think?	*Adaptation*
4. Do teachers deliver the intended dose of 3x12 minute sessions for ten weeks?	*Dosage*
5. What is the rate and scope of participation, at a school, class and pupil level?	*Reach*
6. How well do teachers and pupils engage with the intervention?	*Responsiveness*
7. How different is Stop-and-Think to usual KS2 maths and science teaching?	*Differentiation*
8. What is taking place in the absence of Stop-and-Think?	*Monitoring the control*
9. What outcomes do teachers and pupils perceive to result from Stop-and-Think?	*Perceived impacts*

Research methods

Figure 3 shows how our research activities will address these questions and domains.

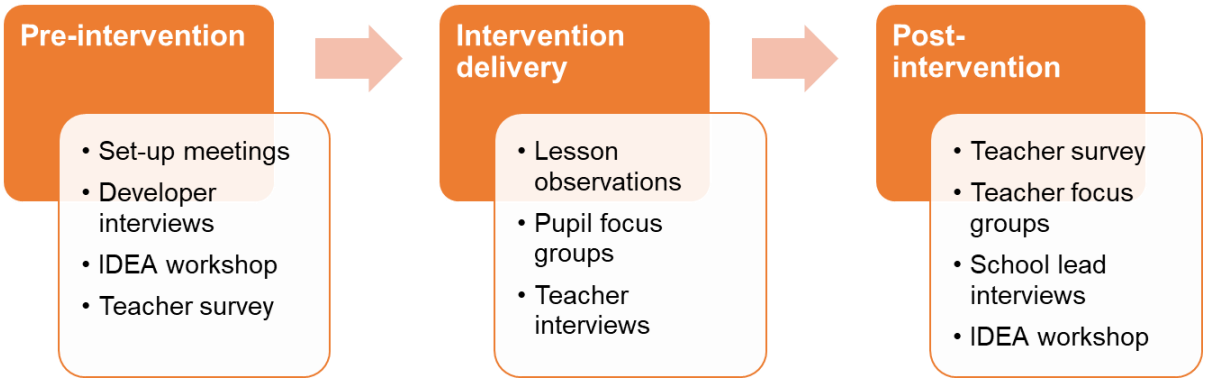
Figure 3: IPE matrix

	Fidelity	Quality	Adaptation	Dosage	Reach	Responsiveness	Differentiation	Monitoring the control	Perceived impacts
Developer interviews	■		■						
Lesson observations	■	■	■			■			
Pupil focus groups						■	■		■
Teacher interviews	■		■		■	■	■		■
Teacher survey	■		■	■	■	■	■		■
Teacher focus groups	■		■			■	■		■
School lead interviews			■			■	■	■	
Post-intervention IDEA workshop	■	■	■		■				
Software data	■			■					

We will conduct the IPE research activities in three phases, as illustrated in Figure 4. The ten-week delivery period means that not all research activities can take place during implementation. We have therefore phased the IPE to ensure activities are timed appropriately to build on one another and minimise burden on schools.

[67] Humphrey, N., Lendrum, A., Ashworth, E., Freason, K., Buck, R. and Kerr, K. (2019) *Implementation and process evaluation (IPE) for interventions in education settings: an introductory handbook*. London: EEF.

Figure 4: IPE phases



Pre-intervention

Pre-intervention IPE activities will identify the nature of and reasons for changes to the intervention from the efficacy trial.

In **set-up meetings and an Intervention Development and Evaluation Analysis (IDEA) workshop** with Birkbeck, BIT and EEF in Autumn 2020, we reviewed the logic model used in the efficacy trial and identified anticipated changes to the software and/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 will inform our examination of the IPE dimensions in Table 3.

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. We will interview the lead developer again in October 2022 to understand any modifications made to the software and delivery plans (and reasons for these), refine and finalise the logic model and explore any findings from their small-scale validation study that should inform our research tools exploring acceptability and feasibility.

We will conduct a **survey of all teachers allocated to treatment and control** immediately prior to delivery to gather quantitative data on business as usual in Autumn Term 2022. This pre-survey will also explore understanding and expectations of the intervention.

During intervention delivery

IPE activities during the delivery period will gather immediate reflections on the acceptability of Stop and Think, fidelity of implementation, reasons for any adaptation and early perceived impacts.

We will **observe lessons** to assess fidelity of implementation and teachers' and pupils' responsiveness. As far as possible, we will aim to observe lessons at a different stage of implementation in each school – for example, in weeks 2, 4, 6 and 8 of implementation – to explore potential differences in these domains (and the domains to be explored in pupil focus groups, described below) over the course of delivery. As well as reporting observations, we will use this data to tailor the prompts and probes used in focus groups and interviews.

We will conduct **pupil focus groups** to explore responsiveness and understand whether and how the programme is encouraging pupils to give slower and more reflective answers. We will explore concrete examples of when pupils feel they have 'stopped and thought' where previously they would not have. During the intervention, pupils will be explicitly told that the programme aims to help them stop and think. Pupil focus groups will therefore look for evidence of explicit understanding as an intended mechanism through which change occurs.

We will **interview teachers** to explore responsiveness, the extent to which the intervention is delivered as intended (including any adaptation/tailoring, and whether this is useful) and how different it is to usual teaching. We will investigate how Stop and Think is incorporated into planning, any support (including technical assistance) required/received and the practicalities of delivering the intervention in different school/class contexts. We will also gather teachers' reflections on early perceived impacts.

Post-intervention

Post-intervention IPE activities will focus on understanding adaptation and variability at scale, as well as exploring perceived impacts.

We will conduct a **survey of all teachers allocated to treatment and control** to gather quantitative data on fidelity, tailoring and adaptation at scale in Summer Term 2023. This post-survey will explore barriers and facilitators to successful delivery and any perceived impacts on intermediate outcomes.

We will conduct **teacher focus groups** with teachers allocated to treatment to explore variation in delivery and identify any reasons for adaptation. The focus group methodology will allow teachers to reflect on their practice more critically by comparing their approach to others.

Each participating school will have a lead member of staff for the Stop and Think intervention and evaluation. We will **interview school leads** to understand business as usual in the year allocated to treatment and monitor the control. We anticipate the school lead will be a subject (maths or science) lead able to speak about both years. We will also ask them about barriers and facilitators to delivery, how the intervention has been incorporated into lesson/curriculum planning and their school's motivations for taking up the intervention.

We will conduct a **post-intervention IDEA workshop** with members of the delivery team to capture their reflections in relation to adaptations to the intervention, reasons behind these changes and possible changes to the logic model. We will also explore technical and substantive support delivered to schools (by regional RAs and the central team) and any challenges and enablers to successful delivery.

Finally, we will analyse **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 by schools). Compliance analysis will be conducted as part of the impact evaluation (see p. 20).

IPE data collection

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

We will conduct interviews by telephone or online according to participant preference. This approach is cost-effective and accommodates teachers' busy schedules. The teacher focus groups will also be online. We will conduct pupil focus groups in-person following lesson observations. Interviews will last 50 minutes, teacher focus groups will last one hour, and pupil focus groups will last max. 45 minutes. All will be audio recorded with participants' permission.

Qualitative data collection will be conducted by researchers from NatCen's Centre for Children and Families. During IPE recruitment and before all IPE data collection activities, the research team will explain to participants that we are independent evaluators (operating separately from the teams at Birkbeck and BIT) and gather informed consent.

The teacher surveys will be administered online and designed to take approximately 15 minutes to complete. We will ask the Stop and Think assistant from BIT to encourage teachers to complete the pre-intervention survey during their set-up visit.

IPE recruitment and sampling

Schools will be purposively sampled to participate in qualitative IPE activities, based on criteria such as proportion of pupils eligible for FSM.[68] Within each school, the following participant groups will be eligible to participate in the IPE: school leads, Year 3 and Year 5 teachers delivering the intervention, and Year 3 and Year 5 pupils receiving the intervention (who have not been withdrawn by their parents – see p. 29). We will select six pupils per school (using sampling criteria such as year group, FSM eligibility, gender) from four schools to participate in pupil focus groups.[69] More details of adult consent and pupil assent to participate in research activities are provided on p. 29.

Table 4 overleaf shows the number of encounters and participants for each IPE activity. Schools will be sampled to achieve diversity in characteristics expected to affect experiences of Stop and Think and/or the way it is incorporated into planning and teaching (e.g. school size, year group receiving Stop and Think, school type).

Analysis

All qualitative interviews and focus groups will be digitally recorded (with permission from participants) and transcribed. All observations will be recorded using detailed fieldnotes.

We will manage and analyse qualitative data using the Framework approach developed by NatCen,[70] which is embedded in in *NVivo 10.* 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 emerge from the data, we will assemble a matrix in which each row represents an individual interview or workshop discussion and each column a theme and any related sub-themes. We will then summarise the qualitative data in the matrix, including illustrative verbatim quotes where appropriate. Once all data has been coded in this matrix, we will move onto analysis. This will involve a phase of 'detection', including studying the elements participants say about a particular phenomenon, listing these and then sorting them thematically. Once we have identified different themes in the data, we will create higher-level categories that work as meaningful conceptual groupings for participants' views and experiences within and across schools.

We will manage and analyse survey and software data using SPSS. Our analysis will explore differences in IPE domains between Year 3 and Year 5, where sample sizes allow.

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

[68] Protocol to be updated with school sampling criteria once agreed with EEF.
[69] Protocol to be updated with pupil sampling criteria once agreed with EEF.
[70] Ritche et al. (2013) *Qualitative Research Practice.* London: Sage.

Table 4: IPE methods overview

Research methods	Data collection methods	Number of encounters and participants	Data analysis methods	Research questions addressed	Implementation/ logic model relevance
Developer interviews	Telephone / online	Encounters: 2 Participants per encounter: 1 **Total participants: 1**	Qualitative	1, 3	Fidelity, adaptation
Lesson observations	Face-to-face	Encounters: 4 Participants per encounter: n/a **Total participants: n/a**	Qualitative	1, 2, 3, 6	Fidelity, quality, adaptation, responsiveness
Pupil focus groups	Face-to-face	Encounters: 4 Participants per encounter: c. 6 **Total participants: c. 24**	Qualitative	6, 7, 9	Responsiveness, differentiation, perceived impacts
Teacher interviews	Telephone / online	Encounters: 8 Participants per encounter: 1 **Total participants: 8**	Qualitative	1, 2, 5, 6, 7, 9	Fidelity, quality, reach, differentiation, perceived impacts
Teacher survey	Online	Encounters: 2 Participants per encounter: c. 200 **Total participants: c. 200**	Quantitative (descriptive)	1, 3, 4, 5, 6, 7, 9	Fidelity, adaptation, dosage, reach, responsiveness, differentiation, perceived impacts
Teacher focus groups	Online	Encounters: 4 Participants per encounter: c. 5 **Total participants: c. 20**	Qualitative	1, 3, 6, 7, 9	Fidelity, adaptation, responsiveness, differentiation, perceived impacts
School lead interviews	Telephone / online	Encounters: 8 Participants per encounter: 1 **Total participants: 8**	Qualitative	3, 6, 7, 8	Adaptation, responsiveness, differentiation, perceived impacts
Post-intervention IDEA workshop	Online	Encounters: 4 Participants per encounter: 1 **Total participants: 4**	Qualitative	1, 2, 3, 5	Fidelity, quality, adaptation, reach
Software data	Via Stop and Think Programme	n/a	Quantitative (descriptive)	1, 4	Fidelity, dosage

Cost evaluation

We will capture delivery costs in the following ways:

- **Costs to schools:**
    - The teacher survey will include questions to establish: whether or how TAs are paid or reimbursed for attending training at each school (in the case of Stop and Think, this would involve attendance at the initial in-school visit); costs of teacher time; any other costs to schools. It will also include questions to explore business as usual costs, i.e. the usual costs of delivering Year 3 and Year 5 maths and science lessons (all schools), and the actual costs of delivering Year 3 and Year 5 maths and science lessons during the trial period (year groups allocated to control).
    - Teacher focus groups will explore any additional time spent by teachers when preparing and delivering maths and science lessons beyond their usual / scheduled time.
- **Costs to the delivery team:**
    - BIT will be asked to complete a pro-forma to indicate the costs of developing the handbook and online videos and the time for Stop and Think assistants to deliver the helpline and ongoing support.

We will also estimate the material costs of equipment required for the intervention, such as computers with whole-class projection or interactive screen facilities and software comparable to Stop and Think. Following EEF's recent guidance on cost evaluation,[71] we will estimate the total cost per school for the intervention as implemented for three consecutive years, and the cost per-pupil-per-school-year (based on the total cost per school divided by the number of pupils per school year expected to benefit from the intervention).

Table 5: Potential resources and evaluation sources for cost evaluation

Category	Item	Evaluation data source(s)	Scope
Personnel for training	Teacher or TA attendance at the initial in-school visit. Within teachers' directed time – no extra cost anticipated. TAs may not be paid for training and meetings – potential extra cost.	Teacher survey (including a question about whether / how TAs are paid or reimbursed for training and meetings)	All schools
Personnel for implementation	Teachers and/or TAs prepare and deliver maths and science lessons, within their usual time – no extra cost anticipated.	Teacher focus groups (to explore how this compares to usual practice)	Teachers allocated to treatment in IPE school sample
Training and programme-level costs	Delivery team develop handbook and online videos	Pro-forma completed by BIT	Delivery team
	Stop and Think assistants deliver helpline and ongoing support	Pro-forma completed by BIT	Delivery team
Facilities, equipment and materials	Computer with whole-class projection or interactive screen facilities	Estimation using market price	Estimation for all schools

[71] EEF (2019) *Cost evaluation guidance for EEF evaluations*. London: EEF.

	Software comparable to Stop and Think		
Business as usual costs	Usual costs of delivering Year 3 and Year 5 maths and science lessons	Teacher survey	All schools
	Actual costs of delivering Year 3 and Year 5 maths and science lessons in control condition during trial period	Teacher survey	All schools

Ethics and registration

Ethical approval

NatCen has a robust ethics governance procedure. Research projects are scrutinised by the NatCen Research Ethics Committee (REC). The REC procedure is designed to provide ethical advice and guidance, and to ensure all research undertaken by NatCen is ethically sound and meets the ethical standards of government and other funders. The process provides reassurance to potential participants and, where relevant, to gatekeepers through whom they are approached.

The NatCen REC has completed a full review of this trial. Ethical approval was granted in November 2020.

Agreement to participate in the trial

Agreement for participation in the trial will be provided by the school headteacher by signing a MoU. As mentioned in the Intervention section (p. 6), schools will be responsible for identifying a Stop and Think school lead, who will liaise with the delivery and evaluation teams. All adult participants (developers, delivery partners, Year 3 and Year 5 teachers, school leads) will be provided with full details about the evaluation so they can give their informed consent to participate.

Schools will give consent *in loco parentis* for Year 3 and Year 5 pupils' participation in the intervention, impact evaluation and IPE. In addition, parents of Year 3 and Year 5 pupils will be provided with full details about the evaluation and given an opportunity to withdraw their child from the impact evaluation and/or the pupil focus groups. We are following this withdrawal approach to parental consent for data collection as all impact evaluation and IPE activities will happen in school, within school hours, and schools will have provided opt-in consent at the point of recruitment.

We will also seek pupil assent for their informed participation in research activities.[72] This will involve the following:

- **A pupil information sheet:** We will prepare a pupil information sheet which explains what pupils will be asked to do during the evaluation, in plain English and an accessible, child-friendly format. This will inform pupils that their participation is voluntary and will have no impact on their academic achievement or what happens to them at school. School leads will be asked to print and share pupil information sheets at least one week prior to research activities, so that pupils have time to familiarise themselves with what they are being asked to do.
- **Verbal assent:** When NatCen researchers and invigilators visit each school, they will verbally explain the evaluation activities in plain English. For pupil focus groups, this will include reminding pupils that they can decide not to participate before or after research activities, and do not need to give a reason for this. For endline testing, invigilators will inform pupils that the tests do not have any effect on their school results, and that pupils can skip questions they do not wish to answer. For all research activities, pupils will be reassured regarding anonymity and

[72] As pupils are under 18, they will not be able to provide informed consent for their participation.

confidentiality of their responses. This will include explicitly stating that their responses will not be shared with their parents/carers, teachers or anyone else at their school. Researchers and invigilators will set aside time to answer pupil questions and address any concerns before starting research activities.

Registration

The trial was registered on 07 May 2021, and the International Standard Randomised Controlled Trial Number (ISRCTN) is ISRCTN12838371.[73] The trial registry will be updated with outcomes at the end of the project.

Data protection

NatCen will collect data during the impact evaluation and IPE. They will store and handle all data securely and confidentially in line with GDPR. The data subjects will be all Year 3 and 5 pupils and staff in participating schools, the developer team at Birkbeck and the delivery team at BIT. Only the research team will have access to data collected as part of the evaluation. This will be monitored through a data security plan, which details all data security procedures to be applied, including names of those who have access rights to respondent confidential data, details of third parties (e.g. transcribers) involved in the project and specific requirements for data destruction. The plan will be updated throughout the project via regular monitoring and internal audits. In the unlikely event of a breach to data security procedures, this will be immediately raised as an Information Security incident. Incidents will be automatically flagged and reviewed immediately by the Principal Investigator and other senior staff in the organisation to agree corrective actions. This will include amendments to the data security plan where required to minimise risk of reoccurrence. You can read more about NatCen's approach to privacy on their privacy policy for this evaluation[74].

In addition, during school recruitment BIT will collect school names, teacher names and teacher contact details. More information about BIT's approach to privacy is available through their privacy policy.[75]

For the duration of the evaluation, NatCen is a data controller who also processes data. This means that NatCen are responsible for deciding the purpose and legal basis for processing data. The legal basis is "legitimate interest".[76] We have carried out a legitimate interests assessment, and have determined that: there is a genuine reason for us to process this data (to evaluate the Stop and Think programme, which will feed into necessary evidence around what works to improve children's maths and science skills), 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. We have assessed the impact our processing might have against the reason for NatCen to conduct this study, and we do not believe the impact would override our interest in any way.

After the evaluation ends, data from the impact evaluation will be stored in the EEF archive. At this point, the EEF will become the data controller. All evaluation data will be securely deleted from NatCen's systems 12 months after submission of data to the EEF data archive and the final edited EEF report (April 2025).

[73] ISRCTN registry, https://doi.org/10.1186/ISRCTN12838371.
[74] NatCen's privacy policy, https://natcen.ac.uk/taking-part/studies-in-field/Stop and Think-evaluation/privacy-notice/
[75] BIT's privacy policy, https://www.bi.team/privacy-policy/
[76] Point (f) of paragraph 1 of Article 6 of the GDPR.

Personnel

The evaluation will be carried out by education and evaluation specialists at NatCen. They bring extensive experience of school-based research, including of other large-scale trials for the EEF.

Table 6: Evaluation team

Name	Project role	NatCen role and team
Mary McKaskill	Principal Investigator and Strategic Lead	Research Director, Children and Families
Helena Takala	IPE Lead and Project Manager	Senior Researcher, Children and Families
Isabel Taylor	Impact Evaluation Lead	Research Director, Evaluation
Enes Duysak	Impact Evaluation Support	Senior Researcher, Evaluation
Migle Aleksejunaite	Impact Evaluation Support	Senior Data Manager, Evaluation
TBC	IPE and testing Support	Researcher, Children and Families
Daniel Phillips	Impact Evaluation QA	Director, Evaluation
Gayle Munro	IPE QA	Deputy Director, Children and Families

Table 7: Delivery team

Name	Project role	BIT role and team
Fionnuala O'Reilly	Delivery Lead	Senior Advisor, Education
Dave Wilson	Delivery Lead	Advisor, Education
Callum O'Mahony	Delivery support	Associate Advisor, Education
Julia Ryle-Hodges	Delivery support	Associate Advisor, Education
Professor Denis Mareschal, University of Birkbeck	Academic Advisor	N/A

Risks

As the Principal Investigator, Mary will be responsible for monitoring and managing risk throughout the project. Table 8 outlines key risks to intervention delivery and evaluation activities and mitigations/contingencies. We will develop this into a detailed risk register during set-up.

Table 8: Key risks

Risk	Likelihood / impact	Mitigations/contingencies
Covid-19 disrupts programme delivery	**Likelihood:** Low **Impact:** High	• Nature of Stop-and Think means delivery should be feasible in case of limits on external visitors • Delivery start date of autumn 2022 allows time for contingency planning in case of school closures
Covid-19 disrupts data collection	**Likelihood:** Low **Impact:** High	• Interviews and focus groups conducted online/by telephone • Observations can be replaced with remote video observation • No testing until summer 2023 • Contingency of teachers administering assessments • EYFSP as a contingency baseline measure for Year 3 pupils if KS1 tests do not go ahead as planned in 2021/22
Recruitment difficulties	**Likelihood:** Low **Impact:** Medium	• Timely recruitment (from November 2021) • Randomisation to allocate treatment and control group in every school to incentivise participation • NatCen to provide clear, concise information on trial requirements • BIT to provide regular updates for progress monitoring
School-level attrition	**Likelihood:** Medium **Impact:** Low	• Trial requirements clearly communicated in recruitment materials/MoU • Experienced field team to schedule testing appointments • Burden minimised through use of single primary outcome • Power calculations account for expected attrition
Pupil-level attrition	**Likelihood:** Low **Impact:** Low	• Mop-up visits to mitigate against pupil absence • No requirement for baseline testing • Power calculations account for expected attrition
Contamination between year groups	**Likelihood:** Low **Impact:** High	• Trial requirements clearly communicated in recruitment materials/MoU • Teachers allocated to control not provided with resources • BIT Stop and Think assistant to reiterate importance of randomisation • Monitoring of control through IPE
Difficulty scheduling fieldwork within required timescales	**Likelihood:** Low **Impact:** Medium	• Data collection points specified in recruitment materials • Online/telephone interviews to accommodate busy schedules • Sufficient resource allocated to arranging visits • Large pool of researchers with relevant skills to draw on if required • Testing in May/June, allowing July as contingency
Low compliance among schools	**Likelihood:** Medium **Impact:** High	• Clearly specified programme • BIT Stop and Think assistant to visit schools to fully explain programme and answer questions

Timeline

Table 9: Timeline

Dates	Activity	Staff responsible/ leading
Jul – Sep 2020	• Initial set up meetings	EEF, Birkbeck, BIT, NatCen
Oct – Dec 2020	• IDEA workshop • Ethical approval • Developer interview	NatCen
	• Recruitment materials developed	BIT, NatCen
May 2021	• Trial registered on ISRCTN	NatCen
Sep – Oct 2021	• Set up meeting with EEF	EEF, Birkbeck, BIT, NatCen
	• Recruitment materials updated	BIT, NatCen
Dec 2021 – Jul 2022	• School recruitment	BIT
March 2022	• Protocol published	NatCen
Jul 2022	• School sample shared with NatCen	BIT
Sep – Oct 2022	• Pupil enumeration • NPD application • Teacher survey • Developer interview	NatCen
Oct 2022	• Randomisation	NatCen
Oct 2022 – Jan 2023	• Scheduling of sessions for teachers • School visits to conduct teacher training	BIT
January 2023	• Statistical Analysis Plan published	NatCen
Jan – May 2023	• Intervention running in schools	BIT
	• Lesson observations • Pupil focus groups • Teacher interviews	NatCen
May – Jul 2023	• Endline testing – Year 3 and Year 5 • Teacher survey • Teacher focus groups • School Lead interviews • Post-intervention IDEA workshop	NatCen

Dates	Activity	Staff responsible/ leading
Aug – Oct 2023	• Analysis and drafting	NatCen
Oct 2023	• Submission of first draft report	NatCen
Apr 2024	• Submission of final edited EEF report • Submission of data to EEF data archive • ISRCTN trial registry updated with results	NatCen

Appendix 1: Changes since the previous EEF evaluation

Table A1: 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.

	Feature	Efficacy to effectiveness stage
		• 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.
Evaluation	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 2: Power calculations for secondary outcome for estimated analytical sample

We have aimed to power this trial to detect and MDES of 0.17 or lower for maths attainment among pupils from disadvantaged backgrounds. This MDES was used as it reflects the effect sizes detected in the efficacy trial.

We selected maths attainment among KS2 pupils from disadvantaged backgrounds as the primary outcome for this trial as it is a key priority for EEF to improve attainment in this subject among this group. As the efficacy trial detected a larger effect size of Stop and Think on maths attainment amongst FSM pupils (0.17) than on science attainment amongst all pupils (0.12), a larger sample of schools and pupils would be required to power this trial to detect a similarly sized effect on science attainment (the secondary outcome). However, a larger effect size on science attainment amongst all pupils (0.20) was found in combined models estimated in Durham University's post-hoc analysis.[77]

Table A2 shows the power calculations used to estimate the minimum sample size required to power the trial to detect an effect of 0.17 for maths attainment both for all pupils and for FSM pupils and 0.12 for science attainment for all pupils only. The number of schools that would be required to power the trial for science attainment among FSM pupils (363) or all pupils (235) is considered to be too large to be feasible for recruitment.

Table A2: Sample sizes required to power trial to detect effect sizes found in efficacy trial

		Maths		Science	
		All pupils	FSM	All pupils	FSM
Minimum Detectable Effect Size (MDES)		0.17	0.17	0.12	0.12
Pre-test/ post-test correlations[78]	level 1 (pupil)	0.635	0.635	0.645	0.645
Intracluster correlations (ICCs) [79]	level 2 (school)	0.07	0.07	0.09	0.09
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15.7	3.5	15.7	3.5
Number of year groups	Treatment	97	165	235	363
	Control	97	165	235	363
	Total	194	330	470	726
Number of schools	Treatment[80]	97	165	235	363
	Control	97	165	235	363
	Total	97	165	235	363
Number of pupils	Treatment	1523	578	3690	1271

[77] Durham University (2020) Re-analysis: Stop and Think: Learning Counterintuitive Concepts. Unpublished manuscript.
[78] Estimates for Year 3 pre-test and post-test correlations are based on the correlation between KS1 and KS2 attainment from FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One London: FFT E.D. Estimates for Year 5 pre-test and post-test correlations are based on the correlation between EYSFP and KS2 attainment found in the efficacy trial. 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.
[79] As found in the efficacy trial. 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.
[80] Due to the randomisation model outlined above, every participating school would have both treatment and control groups.

	Control	1523	578	3690	1271
	Total	3046	1156	7380	2542

However, a sample size of 165 schools at analysis (as estimated to power this trial to detect and MDES of 0.17 or lower for maths attainment among pupils from disadvantaged backgrounds) is also estimated to be sufficient to power the trial for science attainment among all pupils for an effect size of 0.14 and amongst FSM pupils for an effect size of 0.18, as shown in Table A3. These are both lower than the effect size of 0.20 found in the combined models estimated in post-hoc analysis of the efficacy trial. This is therefore used as the target sample size for the number of schools to be included in our analytical sample.

Table A3: Power calculations for minimum sample sizes at analysis

		Maths		Science	
		All pupils	**FSM**	**All pupils**	**FSM**
Minimum Detectable Effect Size (MDES)		0.13	0.17	0.14	0.18
Pre-test/ post-test correlations[81]	level 1 (pupil)	0.635	0.635	0.645	0.645
Intracluster correlations (ICCs) [82]	level 2 (school)	0.07	0.07	0.09	0.09
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		15.7	3.5	15.7	3.5
Number of years per school		2	2	2	2
Number of year groups	Treatment	165	165	165	165
	Control	165	165	165	165
	Total	330	330	330	330
Number of schools	Treatment[83]	165	165	165	165
	Control	165	165	165	165
	Total	165	165	165	165
Number of pupils	Treatment	2587	578	2587	578
	Control	2587	578	2587	578
	Total	5174	1156	5174	1156

Appendix 3: Power calculations for estimated Minimum Detectable Effect Size from the sample of schools, at recruitment

Table A4 shows the power calculations used to estimate the Minimum Detectable Effect Size from the sample of schools and pupils, as at recruitment.

[81] Estimates for Year 3 pre-test and post-test correlations are based on the correlation between KS1 and KS2 attainment from FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One London: FFT E.D. Estimates for Year 5 pre-test and post-test correlations are based on the correlation between EYSFP and KS2 attainment found in the efficacy trial. 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.

[82] As found in the efficacy trial. 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.

[83] Due to the randomisation model outlined above, every participating school would have both treatment and control groups.

The sample size calculations in Table A4 are based on the same assumptions regarding the intra-cluster correlations and pre-test and post-test correlations explained in the Power calculations section of this protocol (see p.14).

Table A4: Power calculations for minimum sample sizes at recruitment

		Maths		Science	
		All pupils	FSM Pupils	All pupils	FSM Pupils
Minimum Detectable Effect Size (MDES)		0.12	0.15	0.14	0.16
Pre-test/ post-test correlations[84]	level 1 (pupil)	0.635	0.635	0.645	0.645
Intracluster correlations (ICCs) [85]	level 2 (school)	0.07	0.07	0.09	0.09
Alpha		0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2
Average cluster size		18.6	4.1	18.6	4.1
Number of years per school		2	2	2	2
Number of year groups	Treatment	181	181	181	181
	Control	181	181	181	181
	Total	362	362	362	362
Number of schools	Treatment	181	181	181	181
	Control[86]	181	181	181	181
	Total	181	181	181	181
Number of pupils	Treatment	3367	741	3367	741
	Controls	3367	741	3367	741
	Total	6734	1482	6734	1482

The power calculations in Table A4 reflect a larger number of schools (181) and pupils (an average of 18.6 pupils tested each in maths and science per year group per school) than in Table 2 or Table A3. This is because the figures in Table A4 reflect the estimated number of schools and pupils to be **recruited to the trial** and the calculations in Tables 2 and A3 are based on the number of schools and pupils estimated to be **included at analysis**. The figures in Table A4 take into account the additional schools we would need to recruit in order to account for expected drop-off of both schools and pupils between recruitment and analysis, which we assume to be at the same level as recorded in the efficacy trial. With efficacy trial-levels of school and pupil drop-out and attrition, we estimate that recruiting 181 schools to this trial would result in 165 schools in the analytical sample. This would allow the trial to be sufficiently powered to estimate the effect of the intervention on the primary outcome: maths attainment among pupils from disadvantaged backgrounds.

[84] Estimates for Year 3 pre-test and post-test correlations are based on the correlation between KS1 and KS2 attainment from FFT Education Datalab (2019) How Attainment Gaps Emerge from Foundation Stage to Key Stage 4: Part One London: FFT E.D. Estimates for Year 5 pre-test and post-test correlations are based on the correlation between EYSFP and KS2 attainment found in the efficacy trial. 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.

[85] As found in the efficacy trial. 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.

[86] Due to the randomisation model outlined above, every participating school would have both treatment and control groups.

Appendix 4: Developing the misconceptions tests

Background

The Stop and Think intervention aims to help pupils adapt to counterintuitive concepts by training them to inhibit their initial response and instead give a more reflective answer. The expected outcomes of the intervention are improved performance on:

- Maths achievement tests (the primary outcome)
- Science achievement tests (the secondary outcome)
- curriculum-appropriate maths and science misconceptions (the intermediate outcome)

Wilkinson et al. (2019)[87] recently developed tests to measure KS2 pupils' common misconceptions; beyond this, there are few examples of studies that have measured common primary-level misconceptions. We will build on Wilkinson et al.'s (2019) approach in two key ways:

- Test length: Wilkinson et al.'s (2019) tests were composed of ten maths and ten science items per year group. We will develop 15-item tests for each subject and year group, as this reduces the likelihood of reaching 'floor' and 'ceiling' effects (see Section 2).
- Test content: Wilkinson et al.'s (2019) tests included items closely linked to content from the Stop and Think intervention. We will develop new items to avoid this overlap, as this reduces the likelihood that correct answers reflect familiarity with the question content, rather than understanding of the underlying concepts.

Test development

We propose developing four tests, as outlined in Table A4 below. Tests will be structured around common KS2 maths and science misconceptions, to be identified through a literature review. Based on our review, we aim to identify five key misconceptions across curriculum domains for each subject. We will also aim to include domains covered by GL Maths and Science Progress Tests, which are being used to measure the primary and secondary outcomes in the effectiveness trial.

We will identify and adapt items to measure each misconception from open source databases such as Diagnostic Questions.[88] We will also review the items from Wilkinson et al.'s (2019) misconception test which did not include material from the Stop and Think intervention.[89] Each item will be presented in multiple-choice format, with the four response options: one correct answer, one 'common misconception' distractor and two other distractors.

To construct a 15-item test for each year group and subject, we propose developing and piloting 30 items per test. For each test, we will develop six pilot items per misconception, with the aim of including three final items per misconception (see Table A4). 15-item tests will allow us to include a greater range of item difficulty, which will reduce the likelihood of reaching floor and ceiling effects (i.e., pupils are less likely to score 0% or 100%). For example, to measure our first misconception in maths (Misconception M1), we would aim to include the following on the final tests:

---

[87] 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'. *Journal of Cognitive Enhancement* 4: 296 – 314.

[88] https://diagnosticquestions.com/. Accessed 12 May 2021.

[89] 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'. *Journal of Cognitive Enhancement* 4: 296 – 314.

- Year 3: three items, ranging from Year 2 to Year 4 difficulty levels
- Year 5: three items, ranging from Year 4 to Year 6 difficulty levels

Our approach to piloting and validating (see below) will allow us to determine the appropriate range of item difficulty for each misconception and for each year group. This test design will also allow us to develop a common scale for Year 3 and Year 5 tests, using Item Response Theory (IRT) modelling. IRT modelling attempts to explain the relationship between latent traits (unobservable attributes) and their manifestations (observed outcomes or performance).[90] In addition to reducing the likelihood of floor and ceiling effects, this approach will allow us to identify any overlaps in misconceptions between year groups (see below for more details).

Table A4: Misconception test structure and number of items

Test	Test structure	Number of pilot items	Number of final items
Year 3 Maths	Misconception M1	6	3
	Misconception M2	6	3
	Misconception M3	6	3
	Misconception M4	6	3
	Misconception M5	6	3
	Total number of items	**30**	**15**
Year 3 Science	Misconception S1	6	3
	Misconception S2	6	3
	Misconception S3	6	3
	Misconception S4	6	3
	Misconception S5	6	3
	Total number of items	**30**	**15**
Year 5 Maths	Misconception M1	6	3
	Misconception M2	6	3
	Misconception M3	6	3
	Misconception M4	6	3
	Misconception M5	6	3
	Total number of items	**30**	**15**
Year 5 Science	Misconception S1	6	3
	Misconception S2	6	3
	Misconception S3	6	3
	Misconception S4	6	3
	Misconception S5	6	3
	Total number of items	**30**	**15**

Pilot fieldwork

We propose a three-stage approach to piloting: a qualitative pre-pilot in five schools, followed by two rounds of validation in 15 schools per round (see Table A5). We plan to conduct all pilot work remotely, to avoid potential Covid-19-related disruptions to fieldwork. NatCen will provide instructions and support to teachers via email and telephone and arrange for a courier to deliver and pick up pilot tests.

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

Table A5: School and pupil sample by pilot fieldwork stage

	Schools	Year 3 pupils	Year 5 pupils
Qualitative pre-pilot	5	25	25
Validation round 1	15	300	300
Validation round 2	15	300	300
Total	**35**	**625**	**625**

We plan to recruit our pilot sample from EEF research school and maths and science networks. We expect that these schools will be familiar with research processes, and so drawing on these networks should facilitate a smooth recruitment and data collection process. We will aim to achieve a diverse sample of schools based on the following criteria:

- region (to include schools across England)
- urban/rural location
- proportion of children eligible for FSM
- school size (large/small schools)

We plan to offer a financial incentive for schools to take part.

Qualitative pre-pilot

In each of the five sampled schools, we will conduct qualitative pre-pilot work to determine construct validity with two teachers (one Year 3, one Year 5) and ten pupils (five Year 3, five Year 5). Within year groups, we will ask a teaching assistant to go through pilot items with pupils using a 'think aloud' protocol. Think aloud methods are commonly used in cognitive test development, and involve asking pupils to verbalise their thought processes as they work through a given scale (see e.g. Johnstone et al 2006). For our misconception tests, this process will indicate how well our items measure counterintuitive concepts, how pupils engage with these concepts when answering items, and whether our items effectively measure intended misconceptions.

We will ask a teaching assistant to work through a think aloud protocol with each pupil while a NatCen researcher observes, takes notes and audio records via video conferencing software. We will ensure that pupils are off-camera during remote observations, to minimise any concerns with video recording children.

We will determine an appropriate number of pilot items for teachers to review with each group of pupils (e.g. ten items per group), to ensure that each observation lasts a maximum of 20 minutes. After each observation, the NatCen researcher will conduct a short interview (30 minutes) with teachers to gain their feedback on the tests, including: suitability of the test content, item difficulty and layout, whether our identified misconceptions are appropriate for the target year groups, and how effectively our items measure each misconception.

Based on the findings from the qualitative pre-pilot, we will revise individual items (e.g. wording, format, response options), test format and structure before the next stage of fieldwork.

Validation rounds 1 and 2

At each round of larger-scale validation, one Year 3 class and one Year 5 class per school in 15 schools will be asked to complete our maths and science misconception tests.

We expect that pupils will need approximately 45 minutes to complete each 30-item test. We will recommend that teachers administer the tests on different days, to avoid pupil fatigue.

Alternatively, if teachers are unable to schedule testing on two different days, we will request that they give a 15-minute break between maths and science tests. Once Year 3 and Year 5 tests have been completed in each school, NatCen will arrange for a courier to collect the completed tests. Tests will then be scanned and four datasets produced ahead of psychometric analysis (see Section A4).

After validation round 1, we will use results from our psychometric analysis to revise or drop items that function poorly (see Section 4). Refined items and tests will then be piloted at validation round 2 in 15 new schools. Finally, we will conduct psychometric analysis using pilot data from validation round 2 to select items for our final misconception tests.

Validity and reliability analysis

Item-level analysis

To assess the validity and reliability of each item, pilot data will be analysed using IRT. Our pilot data will allow us to model two different latent traits:

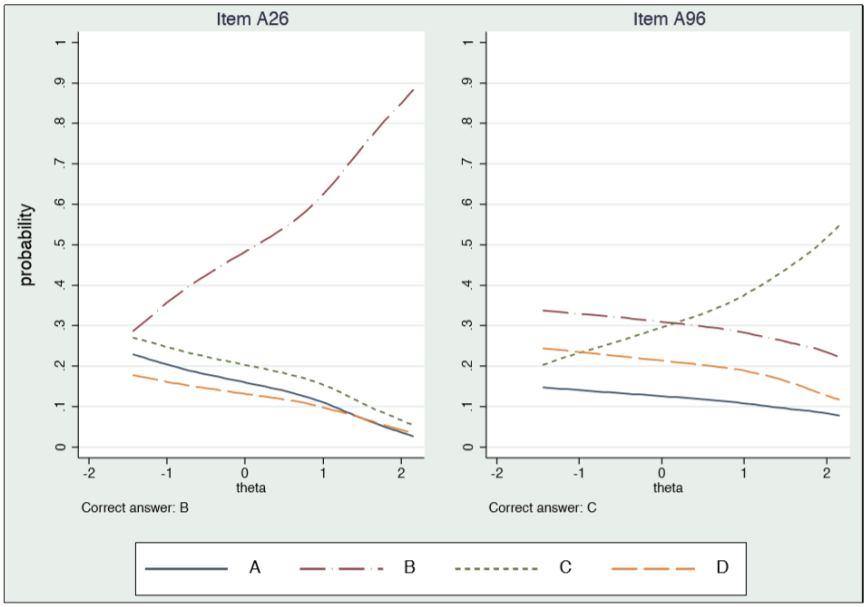
- prevalence of maths/science misconceptions (based on whether pupils choose the misconception option for each item)
- maths/science attainment (based on whether pupils choose the correct answer to each item)

Using two-parameter IRT analysis[91], we will model the probability of choosing a misconception, how this correlates with pupils' overall misconception 'score' and how this relates to their subject attainment.

We expect that lower-ability pupils are more likely to select misconceptions than higher-ability pupils. We will conduct distractor analysis to test this hypothesis for each item. Figure A1 provides an example of a distractor graph, in which pupils' ability[92] (theta) is plotted against the probability of selecting different answer options.

[91] Due to sample size requirements for estimating a three-parameter IRT model, we plan to use a two-parameter model in our analysis. Moreover, a two-parameter model is likely to be more suitable for our purposes because the tests will include distractor options intended to reduce the probability of guessing the correct response to an item. See Azubuike, O. B., Moore, R. and Iyer, P. (2017) *The design and development of cross-county Maths and English tests in Ethiopia, India and Vietnam*. Technical Note 39. Oxford: Young Lives.
[92] Determined by pupils' overall test score.

Figure A1: Example of two distractor graphs (Azubuike, Moore & Iyer 2017)



The distractor graph for Item A96 provides an example of what we would hope to find from an effective misconception item. Assuming that option B is our misconception distractor, the Item A96 graph indicates that:

- pupils are more likely to select the misconception distractor than our 'random' distractors (options A and D)
- as pupil ability increases, they are more likely to select the correct answer (option C).

By contrast, the Item A26 graph suggests that pupils across the ability range are much more likely to select the correct answer (option B) than any of the distractors. This would indicate that the item is not effectively measuring an underlying misconception.

Test and sub-scale reliability

We will assess the reliability of (a) each test as a whole (b) each misconception sub-scale, using an internal consistency measure (Cronbach's alpha). This will provide an indication of the average correlation among all items that make up the test, and each misconception sub-scale. Cronbach's coefficient alpha evaluates the degree to which different items 'pull together' the same content area – and therefore provides a reliable of pupils' underlying trait of interest (in this case, maths and science misconceptions).

Further analyses and item selection

In addition to the analyses described above, we will also examine:

- the overall distribution of misconception scores
- the correlation between maths and science misconceptions
- any overlap in misconceptions between Year 3 and Year 5

Using pilot data from validation round 1, we will use the results of our psychometric analysis to refine the second pilot test for validation round 2. We will then use these pilot data to inform item selection for the final misconception tests, which will be used during endline data collection in the Stop and Think impact evaluation.

We will not use a test-retest approach to assess the reliability of our misconception tests. As we are asking schools to conduct fieldwork remotely, administering tests to the same pupils

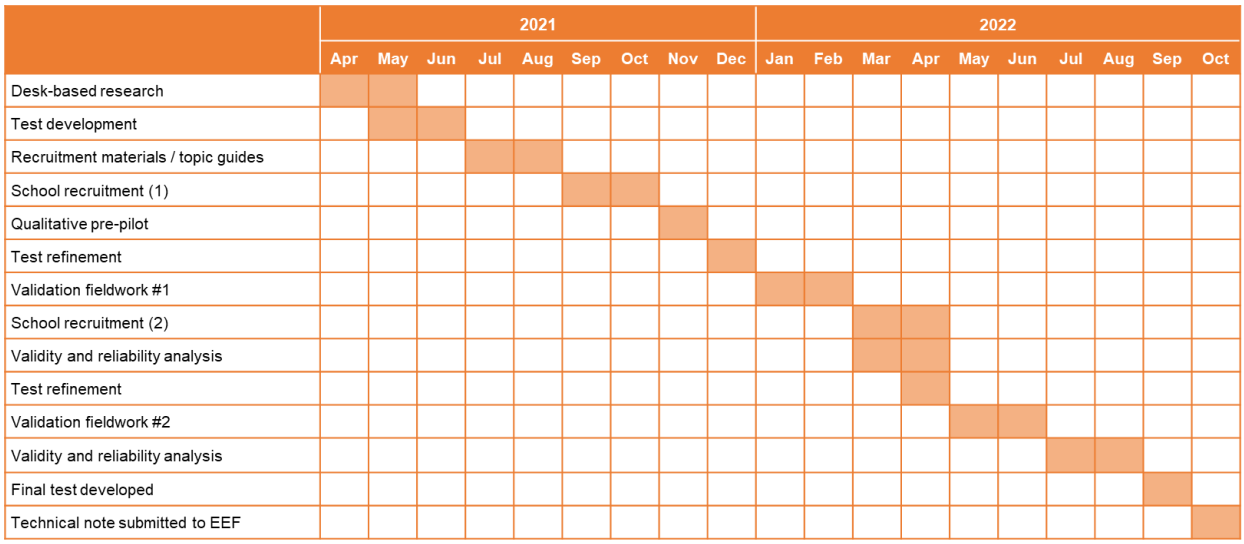
twice within a short period would involve a high research burden for teachers. Instead, our proposed analyses allow a robust approach to determining test validity and reliability while reducing research burden for teachers.

At endline, we anticipate that we will randomise pupils to either complete the maths or science misconception test, to match random allocation to the primary or secondary outcome test. Final data from the impact evaluation (from approximately 6,500 pupils per year group) will allow us to further validate the misconception tests, and to make recommendations regarding the reliability and validity of our items to assess common KS2 maths and science misconceptions in future.

Timeline

Figure A2 provides an overview of our proposed timeline for test development, piloting and analysis.

Figure 2: Timeline for misconceptions test development, piloting and analysis



Appendix 5: School information sheet

Stop and Think: Learning Counterintuitive Concepts

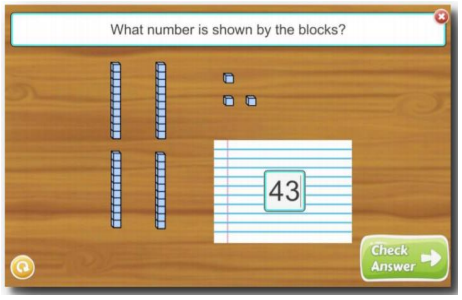
This study is a collaboration between the Education Endowment Foundation (EEF), the Behavioural Insights Team (BIT), the National Centre for Social Research (NatCen) and the Centre for Educational Neuroscience at Birkbeck, University of London.

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).



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!

Appendix 6: Parent information sheet

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: _____

Appendix 7: Memorandum of Understanding

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).

Appendix 8: Privacy notice

Stop and Think: Learning Counterintuitive Concepts

Privacy Notice

In line with the EU 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).

You can find out more about NatCen at www.natcen.ac.uk.

You can find out more about the EEF at www.educationendowmentfoundation.org.

Who will access personal data?

NatCen are carrying out this evaluation and named individuals on the NatCen research team will have access to:

- school and teachers' names and contact details
- audio recordings from pupil focus groups, teacher focus groups and teacher interviews
- teacher survey responses
- sample files provided by schools containing pupils' names, dates of birth, gender, free school meal (FSM) eligibility and unique pupil numbers (UPNs) and the class they are in at school
- pupil attainment and assessment data

McGowan Transcriptions (www.mcgowantranscriptions.co.uk) is the transcription service NatCen use to transcribe our interview data. They will have access to recordings and transcripts from all interviews. 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 use data from pupil assessments and 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 data will be pseudonymised before being analysed.

Information and opinions gathered from pupil focus groups, teacher focus groups, teacher 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 data will be treated with the strictest confidence – no schools, teachers or pupils will be identified in any report arising from the research.

NatCen will share data from the impact evaluation (including information from the NPD) with academics at Birkbeck University for the purposes of further scientific study, 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. Other research teams might match this data to other information from the NPD and other administrative data in the future.

NatCen will securely delete personal information about participants no more than six months after the evaluation is finished (by September 2024 at the latest).

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 couldn’t 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, 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/.

You also 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 email us via StopAndThinkEvaluation@natcen.ac.uk or call 020 7549 7190 during office hours.

Contact information

If you have any questions about the evaluation, including how personal information will be processed, please contact the evaluation team at StopAndThinkEvaluation@natcen.ac.uk.