



Stop and Think: Learning Counterintuitive Concepts

Evaluation Report

September 2019

Palak Roy, Simon Rutt, Claire Easton, David Sims, Sally
Bradshaw and Stephen McNamara

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education. Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

This project was funded as part of the Education and Neuroscience scheme, which was jointly funded by Wellcome and Education Endowment Foundation and launched in January 2014. The aim of the scheme was to provide funding for collaborative projects between educators and neuroscientists to develop evidence-based interventions for use in the classroom, or to rigorously test existing tools and practices.

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



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



0207 802 1653



jonathan.kay@eefoundation.org.uk



www.educationendowmentfoundation.org.uk

Contents

About the evaluator	3
Executive summary	4
Introduction	6
Methods.....	14
Impact evaluation	29
Implementation and process evaluation	44
Conclusion.....	53
References	55
Appendix A: EEF cost rating.....	59
Appendix B: Security classification of trial findings.....	60
Appendix C: Main trial information and consent.....	62
Appendix D: Memorandum of Understanding	67
Appendix E: Theory of Change (ToC)* for the External Evaluation of Learning Counterintuitive Concepts	68
Appendix F: Randomisation code (SPSS syntax)	69
Appendix G: Stop and Think and SEE+ teacher survey	87
Appendix H: Histograms of Prior Attainment, EYFSP scores.....	97
Appendix I: Distribution of outcomes measures (PTM8, PTM10, PTS8 and PTS10) by randomisation groups	99
Appendix J: Histograms of secondary outcome measures from the Chimeric Animal Stroop task	101
Appendix K: Number of Stop and Think sessions (compliance measure)	102

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER). The principal investigator for this trial was Simon Rutt, Head of Centre for Statistics. Palak Roy, Senior Trials manager, led the evaluation team.

Contact details:

National Foundation for Educational Research
The Mere
Upton Park
Slough
Berkshire
SL1 2DQ
p: 01753 637385
e: p.roy@nfer.ac.uk

Acknowledgements

We are grateful to all the primary schools that participated in the pilot as well as in the main study. We are especially thankful to the classroom teachers, senior leaders, and headteachers who participated in NFER case study visits and telephone interviews. We are also thankful to Tom Dickinson, Max Falinski, Kathryn Hurd and Shalini Sharma from NFER's research operations for their endless efforts to maintain ongoing relationships with the schools; Sagina Khan and Emma Hawkins from NFER's Research Department for their administrative support throughout the evaluation; and Afrah Dirie for her support with the impact evaluation analysis. Huge thanks are also due to Dr Ben Styles, Head of NFER's Education Trials Unit for his guidance and advice throughout the trial. We also appreciate the support of the team at the Education Endowment Foundation, including Eleanor Stringer, Camilla Nevill, Dr Anneka Dawson, Dr Florentina Taylor and Amy Clark for their guidance throughout this evaluation.

We would like to thank the delivery team at Birkbeck College and UCL Institute of Education for their support throughout the evaluation. The delivery team also contributed to the introduction section specifically by providing the rationale for the intervention from a neuroscience perspective.

Executive summary

The project

The Learning Counterintuitive Concepts project aimed to improve science and maths attainment for Year 3 and Year 5 pupils (aged 7–10) using an intervention called Stop and Think. When learning new concepts in science and maths, pupils must be able to inhibit prior contradictory knowledge and misconceptions to acquire new knowledge successfully. Stop and Think is a computer-assisted learning activity that aims to improve learner's ability to adapt to counterintuitive concepts by training them to inhibit their initial response and instead, give a slower and more reflective answer.

The intervention was developed at the Centre for Educational Neuroscience, by a team from Birkbeck University of London and the UCL Institute of Education, and evaluated as part of a round focused on neuroscience co-funded by The Wellcome Trust and the EEF. The intervention, derived from cognitive neuroscience principles, was delivered by teachers to the whole class and consisted of thirty sessions being delivered for a maximum of 15 minutes, three times a week, for ten weeks at the start of maths or science lessons.

This project was a randomised controlled trial. Eighty nine schools were randomly allocated to have either Year 3 or Year 5 as their intervention year receiving Stop and Think, with the other year group acting as one of the two control groups. Half of the control years were 'business as usual' that continued with normal classroom practice, and half received a computer programme to support social/emotional skills as an active control condition. This meant that we could measure specific effects of the Stop and Think intervention beyond additional engagement and motivation caused by the novelty of playing a computer game. The primary outcomes were maths and science attainment and the project also looked at a general measure of inhibitory control as a secondary outcome.

Key conclusions

1. Children in the intervention group made the equivalent of one additional month progress in maths and two additional months' progress in science, on average, compared to children in the control group. The maths result is not statistically significant. This means that the statistical evidence does not meet the threshold set by the evaluator to conclude that the true impact was non-zero. These results have a high security rating.
2. The use of two primary outcomes increases the risk that a false positive result may be found through chance. The mixed results between the two outcomes mean that the evaluator is unable to conclude that the programme is effective at raising attainment outcomes.
3. The project found no evidence that the Stop and Think programme had an impact on pupils' general inhibitory control.
4. A majority of teachers thought that Stop and Think had a positive impact on the mathematical and science abilities of the pupils in their class. Other impacts of using the programme included pupils taking time to consider their response before answering questions, enhanced confidence and improving engagement in learning.
5. The majority of teachers did not endorse the roll out of the programme in its current form to other schools. The most common reasons given were the difficulty in fitting delivery into the school day, software problems, pupil engagement, the accuracy of content, quality of animation and some of the content being too easy.

EEF security rating

This trial was a well-designed efficacy trial to test whether the Stop and Think intervention can work under developer-led conditions in a number of schools. Baseline imbalance for the analysed groups suggested that the pupils in the intervention group were similar to those in the control and control-plus groups in terms of their prior attainment. Due to pupils having left the school or being absent on the day of testing, 17% of pupils from maths and 16% pupils from science were not included in the final analysis. The trial security rating was therefore reduced to four padlocks.

Additional findings

The combined effect size (across Year 3 and Year 5) in maths and combined effect size in science were the joint primary outcomes for this trial. The decision to use two primary outcomes increases the risk that a false positive result is found through chance—this risk increases from 5% to 9.75%. The mixed results between the two outcomes mean that the independent evaluation team, therefore, are unable to conclude that the intervention had a positive impact. While we

acknowledge that having two statistically significant results is a more conservative approach and therefore a higher bar to set for any evaluation, these results will need to be considered alongside other findings from the impact and process evaluations.

Looking at attainment in science and maths for each individual year group, the analysis found that the intervention had a positive effect on Year 5 pupils' science attainment. The project did not find any statistically significant evidence of impact on Year 5 pupils' maths, Year 3 pupils' maths, or Year 3 pupils' science attainment. There was no evidence that Stop and Think had an effect on pupils' general inhibitory function development (measured by the Chimeric Stroop task).

Looking at the combined effect size (across Year 3 and Year 5), children who received Stop and Think made more progress than children in the active control group. These results were statistically significant for both maths and science. The results demonstrate that the Stop and Think programme had an impact on pupils' maths and science attainment over and above a similar computer programme.

There were mixed results for pupils who were eligible to receive free school meals (FSM) any point in the previous six years. For Year 3 and Year 5 maths, and Year 5 science, FSM pupils made additional progress, on average, compared to the control group. This was not the case for Year 3 science where we found no additional progress compared to the control group. However, the study was not powered to measure an effect for FSM pupils and the effects were not significant.



Most teachers felt well supported and indicated that their school did not require any additional resources to run the programme. However, over half of teachers reported experiencing issues using the software, which caused delays and impeded the smooth running of the sessions. Teachers also suggested that Stop and Think could be improved if it offered teachers more control of the topics that came up so that they could use it to refer to topics already covered by their class.

Costs

The average cost of Stop and Think was £5.76 per pupil per year when averaged over 3 Years. This estimate is based on the delivery of the intervention to one year group. It is estimated on the basis of the programme software being free, and includes costs of the initial training and ongoing support from Birkbeck provided in this trial for the first year only. The assumption is that schools could use the handbook for the subsequent two years without training. This estimate does not include costs associated with staff time such as training and preparation. Schools estimated, on average, that the time involved in preparing for and setting up Stop and Think was less than five minutes, and the average length of the one-off training at the start of the year was less than thirty minutes.

Impact

Table 1: Summary of impact on primary outcomes of maths and science (GL test scores)

Outcome/ Group	Effect size (95% confidence Interval)	Estimated months' progress	No. of pupils	P value	EEF security rating	EEF cost rating
Maths (Year 3 and Year 5 combined) vs control	0.09 (-0.01, 0.19)	1	2,702	0.087		£ £ £ £ £
Science (Year 3 and Year 5 combined) vs control	0.12 (0.02, 0.22)	2	2,735	0.018		£ £ £ £ £

Introduction

Learning of Counterintuitive Concepts (also known as UnLocke) is a research project that focuses on understanding the learning of counterintuitive concepts in science and mathematics education through a behavioural and neuroimaging¹ study of Year 3 and Year 5 primary school children participating in subject-specific inhibitory control training (in comparison to social skills training or lessons as usual). Children's ability to learn science and maths concepts is limited by their ability to inhibit perceptual evidence (what they see, feel, or hear) or pre-existing beliefs (Babai *et al.*, 2015; Rouselle *et al.*, 2004; Borst *et al.*, 2013; Linzanni *et al.*, 2015; Lubin *et al.*, 2013; Vosniadou *et al.*, 2018a; Brault Foisy *et al.*, 2015; Masson *et al.*, 2014; Stavy and Babai, 2010).

For example, children learn that the world is round, whereas there is no direct visual evidence to support this idea as the horizon looks flat—counterintuitive concepts (Allen, 2014). Many mistakes in maths and science are made because children have a tendency to answer with an intuitive response (Vosniadou *et al.*, 2018b). The intervention in this study aims to train children in a cognitive strategy meant to make them reflect, or 'stop and think', about science and maths problems before answering. 'Stop and Think' is a computerised learning activity that uses content based on the maths and science curriculum of Year 3 and Year 5 children in England. This was the intervention for the evaluation. 'See+' uses a similar computerised platform with content based on the Personal, Social and Health Education (PSHE) curriculum (and specifically not maths or science related). This was offered to the 'control-plus' group. The main difference between the two tasks is the domain that they target. Whereas Stop and Think ultimately aims to improve academic performance in maths and sciences, See+ aims to help children become more proficient at analysing and understanding different forms of social interactions. The aim of this study was to assess whether Stop and Think improves science and maths performance in primary-school-age children delivered via a computerised game. The See+ computer programme was introduced as 'control-plus' to discern effects of the intervention from effects of using a novel computer programme.

Conceptual understanding of reasoning is relevant to education—pupils sometimes demonstrate misconceptions based on faulty thinking (Mareschal, 2016; Vosniadou *et al.*, 2018b). This arises when pupils are asked to reason about counterintuitive concepts, especially in maths and science where they find it difficult to inhibit or suppress their intuitive reasoning. Mareschal (2016) defines counterintuitive concepts as follows:

A key element of learning any new concepts is the need to overcome strongly held prior beliefs about a domain before new knowledge can be effectively assimilated. Thus, a major challenge in mathematics and science education is the need for children to inhibit pre-existing beliefs or superficial perception in order to engage in acquiring and applying new and counterintuitive knowledge.

Thus any pupil aiming to acquire 'new' concepts in science and mathematics needs to overcome the strong pull of existing beliefs.

An example, would be where a pupil believes that the cell size of an elephant is larger than that of a mouse but learns in science lessons that the cell size of both animals is the same. A further example is that, when pupils are taught about negative numbers in maths, they are likely to make the mistake of thinking that -5 is larger than -1 (Bofferding, 2019). In science, pupils tend to think that the sun appears to move in the sky rather than the earth revolving around the sun. UnLocke observes that

misconceptions are particularly common in maths and science. In science education, it can be a real challenge for children to acquire knowledge that goes beyond popular beliefs or perceptions, while in maths children need to go beyond the perceptively obvious solutions to uncover formal logical solutions to a problem.²

The Stop and Think intervention was derived from cognitive psychology and neuroscience research, as noted on the developer's website (see footnote). Specifically, the rationale for this intervention was informed and underpinned by a theoretical understanding of the ways in which people reason and make decisions. Evans (2003) posited that two competing cognitive systems underlie reasoning: the heuristic system, which is evolutionarily, old, fast operating,

¹ Note that this is not part of this evaluation. This relates to additional research activities undertaken by the researchers at Birkbeck College.

² <http://unlocke.org/neuroscience.html>

automatic, and parallel (sometime called ‘System 1’) and the analytic system, which is slow operating, rule-based, and sequential in nature (‘System 2’). The analytic system underlies abstract logical reasoning and hypothetical thinking but it is limited by how much we can keep in mind at any point (working memory capacity). A defining property of the dual process model of reasoning is that the analytic system is able to inhibit and override the heuristic system so that individuals can think things through and successfully carry out logical tasks instead of giving an automatic, immediate, incorrect response (Evans, 2003; Houdé and Tzourio-Mazoyer, 2003). Neuroimaging work on logical and scientific reasoning in adults has consistently shown that the inhibition of pre-existing beliefs, misleading perceptual-biases, and intuitive heuristics is associated with the activation of the anterior cingulate cortex (ACC) and the prefrontal cortex, notably the inferior frontal cortex (IFG) and dorsolateral prefrontal cortex (DLPFC) (Borst *et al.*, 2013; Fugelsang and Dunbar, 2005; Dunbar *et al.*, 2007; Goel and Dolan, 2003; Masson *et al.*, 2014; Prado and Noveck, 2007; Stavy and Babai, 2010). Critically, Houdé *et al.* (2000) provided evidence of a switch, after a brief training, from the heuristic system to the analytic system, with an associated shift from the recruitment of posterior brain regions to the recruitment of a left fronto-parietal brain network.

An important aspect of the Stop and Think intervention is that it is embedded in maths and science reasoning. Although there have been several training programmes targeting executive functions in young children, these have had limited success at generalisation or transfer to other domains (working memory training: Shipstead *et al.*, 2012; inhibition training: Thorell *et al.*, 2009; attention training: Kerns *et al.*, 1999; Wass *et al.*, 2012). Standard information processing approaches to cognition (that abstract away from neural processes) represent processes as encapsulated modules (for example, attention module, working memory module). However, it has been argued that in reality, control of knowledge within neural networks is embedded within particular domains of knowledge (McClelland and Rogers, 2003). Therefore, training domain general skills (such as a putative general working memory capacity) may not have as much impact on the control of knowledge as training within a target domain. There is therefore a need to develop inhibition-training programmes that go beyond the current domain-general approaches. This is a key insight that underpinned the Stop and Think intervention, grounded in understanding of neural information processing (McClelland and Rogers, 2003). A few interventions have implemented cognitive control training within the classroom environment or within maths and science (Diamond and Lee, 2011; Kusché and Greenberg, 1994; Riggs *et al.*, 2006; Stavy and Tirosh, 2000). Results show long-term effects (Riggs *et al.*, 2006) and more generalizable benefits when the training is embedded within the curriculum than when it is not (Diamond and Lee, 2011). Thus, using an embedded approach where training takes place *within* the maths and science curriculum, the Stop and Think intervention aims to train pupils to engage their System 2 analytic reasoning while at the same time inhibiting their System 1 reasoning thereby enabling them to give more considered, reflective, and correct responses to questions.

The intervention is relevant to improving maths and science education, which is high on the government’s education policy agenda to promote STEM (science, technology, engineering, and maths). STEM expertise is considered key to improving the U.K.’s economic growth and productivity. International surveys reveal that the U.K. faces challenges in improving young peoples’ maths and science skills. For example, PISA results (2015) showed that 15-year-olds in the U.K. ranked outside of the top ten of the 72 countries which took the assessments (15th in science and 27th in maths; OECD, 2018). Kuczera *et al.* (2016) reported that over a quarter of young people aged 16–19 in England had low numeracy (below level 2) skills, placing England 22nd out of 23 countries. The CBI/Pearson Education and Skills Annual Report, 2017 (CBI, 2017) found that a majority of the employers surveyed said that STEM skills should have central importance in primary and secondary education. Employers valued problem-solving skills, resilience, and communication as well as literacy and numeracy skills when recruiting school and college leavers. As noted above, analytic reasoning (System 2) supports scientific and numeric problem-solving. The government acknowledges the issue and is attempting to address it in several ways. As noted in the government’s Productivity Plan (House of Commons, 2015), the government has introduced new and more rigorous GCSE and A-levels in maths and science and aims to train an additional 17,500 teachers in STEM subjects. The government’s Industrial Strategy (HM Government, 2017) highlights a range of interventions designed to drive up the study of maths including an expansion of the Teaching for Mastery maths programme which aims to reach 11,000 primary and secondary schools by 2023.

The rationale for evaluating this particular intervention was to explore how insights from neuroscience can be used to improve education. Evidence from neuroscience research supports the hypothesis that inhibitory control is necessary to develop the reasoning skills required in maths and science (Babai *et al.*, 2015; Brault Foisy *et al.*, 2015; Masson *et al.*, 2014; Vosniadou *et al.*, 2018b). Studies of interventions designed to improve such ‘executive function’ skills have shown improvements on outcomes like working memory, but have often failed to show an impact on broader attainment

measures (Diamond and Ling, 2016). Emerging neuroscience research suggests that the inhibition needs to happen in the networks that are specific to the skills being developed, thus the need for exercises to be related to specific subject knowledge (Botvinick and Cohen, 2014; McClelland and Rogers, 2003; O'Reilly *et al.*, 2010). As a result, the focus of this project was on developing exercises that are more closely related to the curriculum areas of science and maths. The aim was to test whether practising these skills leads to improvements in attainment in subject tests.

The evaluation was set up in the autumn term of 2015 and a trial protocol was published during this phase (NFER, 2016). The evaluation had two phases: (1) an 18-month development and pilot phase (January 2016–July 2017) and (2) a randomised controlled trial phase (main trial, January 2017–April 2018). The purpose of the development part of the pilot was (1) to develop and test the intervention materials with the pilot schools (Birkbeck College), (2) assess the suitability of three trial groups for implementation and feasibility (Birkbeck College and NFER) and (3) to determine the best way of delivering the computerised intervention: one-to-one individualised or whole-class (Birkbeck College and NFER). NFER conducted a small-scale process evaluation comprising interviews with staff in two pilot schools using Stop and Think and one school using See+ to explore the feasibility and scalability of the intervention. Following this, NFER shared a summary of the findings with the EEF and Birkbeck College and the theory of change (TOC) model for the intervention was devised. The recruitment for the main trial commenced in January 2017 prior to completion of the pilot.

The findings from this stage suggested that having two groups (intervention and control-plus or control classes) within one school was practical. We found this was not an issue as the intervention, control-plus, or control classes were in different year groups. As a result of the pilot study and findings from the evaluator and the developers' own experience of implementation, the main trial delivered the intervention (Stop and Think) and control-plus software (See+) in a whole-class setting rather than delivered to pupils individually within a class setting, for example, one computer per pupil. The implementation and process evaluation methods section includes further details on the findings from this phase.

Intervention

The intervention, Stop and Think, in this study aims to train children in a cognitive strategy meant to make them reflect, or 'stop and think', about science and maths problems before answering. Stop and Think is a computerised learning activity that uses content based on the maths and science curriculum of Year 3 and 5 children in England. See+ uses a similar computerised platform with content based on the Personal, Social and Health Education (PSHE) curriculum. The main differences between the two approaches are the domain that they target (social versus maths or science) and the fact that See+ was not designed to train the use of inhibitory control skills.

Why—rationale/theory

Evidence from neuroscience supports the hypothesis that inhibitory control is necessary to develop the reasoning skills required in maths and science. Stop and Think draws on work which suggests that being trained in inhibition control engages parts of the brain required for logical thinking and for learning new concepts in maths and science. When learning new concepts in science and maths, pupils must be able to inhibit prior contradictory knowledge to successfully acquire new knowledge (although conceptual change is relevant for a range of subjects, misconceptions are particularly common in maths and science). It is thought that using a computer programme will engage pupils during maths and/or science lessons in trying to solve problems that will enable them to practice counterintuitive learning and reasoning skills by engaging the pre-frontal cortex.

Who—recipients

The delivery of the programme was facilitated by teachers and delivered to classes of Year 3 or Year 5 pupils (7–10 Year olds) during the autumn term in the academic Year 2017/18.

What—materials

The Centre for Educational Neuroscience, a collaboration between Birkbeck, the Institute of Education, and University College London, developed a computer-assisted learning activity in 2016–2017 to train a pupil's ability to control such interferences. The computer-based learning activity is designed to help children in Years 3 and 5 stop and think before tackling problems in science and maths. A friendly character, named Andy, poses questions to three virtual game-show

contestants who demonstrate correct and incorrect thinking. Children complete various tasks as if they are taking part in the game-show.

The software programme was set up by teachers at the start of each session. Teachers decided how to facilitate the sessions, and how the pupils interacted with the software to input the 'answer'.

Who—implementers

Teachers facilitated the sessions with pupils in a whole-class format. The questions are given in a pre-recorded audio within the Stop and Think programme. The questions are posed by the main character in the programme ('Andy'). There was the option of teachers reading the text out as well but this was not necessary. Teachers facilitated the sessions with pupils.

How—mode of delivery

Birkbeck recommended delivering the session for a maximum of 15 minutes three times a week at the start of a maths or science lesson. Sessions were initiated and facilitated by teachers who guided pupils through the task. The computer-based learning activity was set up like a game show in which the host, Andy, posed questions to the pupils and three virtual game show contestants. These characters took pupils through the maths and science tasks, providing prompts and demonstrating the correct way to think about these concepts. Andy and the game show contestants offered different levels of support dependent on pupils' responses.

The questions posed by Andy were considered by the whole class. The teacher facilitated the process of agreeing a group response to the questions and entering an answer which could be right or wrong. If the answer entered was wrong, the programme gave prompts to stimulate pupils' thinking which would enable them collectively to get to the right answer.

Where—setting

Sessions took place in class, adopting a whole-class approach mostly using an interactive whiteboard. The Stop and Think programme questions were loaded on to a laptop and projected on to a whiteboard and the teacher and pupils agreed an answer to each question which was then entered in the programme. The teacher and pupils then viewed the programme's response which indicated whether the answer they had given was right or wrong. If it was right they moved on to the next question. If the answer was wrong, they then considered the feedback provided by the programme and entered a different answer.

When and how much—dosage

The Stop and Think programme was to be delivered at the start of maths or science lessons. Each session was made up of multiple subtasks relating to one maths topic and one science topic based on age-relevant content from the National Curriculum. Topics were delivered in a fixed order for consistency across schools for the purpose of the evaluation. Each session lasted for a maximum of 15 minutes and was delivered three times a week, for ten weeks (30 sessions in total). The software had a built-in 12-minute timeout function to try to ensure sessions only replaced approximately 15 minutes of lessons (including set-up time) and ensure consistency across schools.

Tailoring

Teachers were given an opportunity familiarise themselves with the software and were advised to deliver the sessions as they wished. This meant that teachers had some flexibility in delivery to be responsive to their school context.

How well—planning

Strategies to maximise implementation effectiveness included attendance at in-school training sessions and having a named Birkbeck researcher linked to each school for support as needed. Written guidance was also given to each school. The training involved one face-to-face familiarisation session where the Birkbeck researcher installed the software on the class computer, demonstrated how to use the programme, and explained delivery method/s. This

process was responsive to individual teacher's needs and questions and not time-limited. The researcher was also available to deal with any follow-up queries from the teacher by telephone or email.

Birkbeck noted that teachers did not need to be formally trained to use the programme; they could read the teacher guide and run the intervention. Teachers were also given a 'teacher information pack' which contained information about the evidence and theory underpinning the programme, about the programme itself and frequently asked questions (FAQs).

Implementation of Stop and Think

The sessions we observed were whole-class and overseen by a teacher who invited the class to give an answer agreed by the majority of pupils to each question asked by the programme which was then inputted to the programme. This was repeated for each question. In these sessions, teachers sometimes asked pupils to explain why they had given a particular answer but did not influence pupils' answers. The teachers or a pupil then entered the answer in the programme. The teacher survey found that 35 of the 61 teacher respondents indicated that they delivered the session at the start of maths or science lessons and a further 18 teachers said that this happened 'sometimes'. We observed sessions delivered in schools where classes included two year groups of pupils and the teacher targeted the session on the year group eligible for inclusion in the programme. Most of the teachers surveyed (50 out of 61 responding to the survey) said that the training was suitable in preparing them to use the Stop and Think programme.

The research assistant at Birkbeck College communicated incorrect group assignment to two schools. Due to this communication error, the year groups that were randomised to the control group—those that were not meant to receive any computer programme—implemented the intervention. This is considered as an administrative error and not a contamination as the schools continued the practice as they were asked. These schools, despite delivering the intervention to year groups that were randomised to the control group, are analysed as randomised to fulfil an intention-to-treat (ITT) analysis. A further two schools corrected their form-entry set-up information after randomisation, which means they belonged to incorrect randomisation strata. Again, we analysed these schools as randomised to fulfil ITT analysis.

See+—control-plus group

See+ was offered to schools as part of the research project. Classes using See+ acted as the 'control-plus' group, allowing the evaluation to examine whether improvement in inhibitory control and academic activity was a result of using the Stop and Think programme specifically, or the result of having a novel computer-based activity at the start of lessons more generally. In order to rule out the latter, Birkbeck College developed a programme that did not include any content from Stop and Think; the only similarity between the intervention and the control-plus groups was having access to a novel computer-based programme.

Why—rationale/theory

See+ is a socio-emotional skills learning tool developed by the team at Birkbeck College for pupils to use at the start of lessons that are not maths or science, usually Personal, Social and Health Education (PSHE) lessons. See+ learning sessions were developed for the purpose of this research project and piloted prior to the main trial. The See+ programme aligns with the PSHE national curriculum and with the Social Emotional Aspects of Learning (SEAL) and the Social, Emotional Regulation and Transactional Support (SCERTS) curricula.

The capability to understand other people's intentions, emotional states, and beliefs is known as 'social-emotional cognition'. See+ learning activities aim to develop pupils' social-emotional cognition by raising their awareness of other people's perspectives about respect, fairness, equality, and social behaviour.

Who—recipients

The delivery of the programme was facilitated by teachers and delivered to classes of Year 3 or Year 5 pupils in the same school during autumn term in the academic year 2017/2018. The design of the trial meant that one year group in a school received the Stop and Think programme and the other year group either received See+ or continued with normal classroom practice.

What—materials

The See+ learning activity in PSHE lessons was set up as a computerised animated story with virtual characters engaging in different social scenarios. Each session covered one social-emotional learning theme. Within each session, all pupils engaged in learning tasks presented as a sequence of three sub-tasks: observation, followed by comprehension, followed by reflection.

Who—implementers

Teachers were trained in how to use the software and were advised to limit their support to reading the task text only: no prompts were to be given by the teacher regarding the interpretation of the social scenario presented on screen. The voice of the main character in the programme ('Andy') was audible—he spoke his questions out. There was the option of teachers reading the text out as well but this was not necessary. Teachers facilitated the sessions with pupils.

How—mode of delivery

See+ was delivered as a whole-class activity. Each task in See+ included prompts to help teachers navigate pupils through the session. Pupils watched the animation and then reflected on what they had seen by answering a series of comprehension and reflection questions about the characters' interaction and socio-emotional state. Pupils had the opportunity to discuss and think about how the characters might have resolved any difficulties or dilemmas they experienced. The sessions were delivered through oral whole-class discussions.

Where—setting

Sessions took place in class, adopting a whole-class approach. The See+ programme animations and questions were projected onto a whiteboard and the teacher and pupils discussed each question and agreed an answer which was then entered. The process was repeated throughout the session.

When and how much—dosage

Each session lasted for a maximum of 15 minutes and was delivered three times a week for ten weeks (30 sessions in total).

Tailoring

Teachers were given the opportunity to familiarise themselves with the software and were advised to deliver the sessions as they wished at the start of lessons, excluding maths or science. This meant that teachers had some flexibility to be responsive to their context and fit in sessions with the school timetable.

How well planned

Strategies to maximise implementation effectiveness included attendance at in-school training sessions and having a named Birkbeck researcher linked to each school for support as needed. Written guidance was also given to each school.

Implementation of See+

The See+ sessions we observed were whole-class and overseen by a teacher who invited the class to give an answer to the questions posed by the programme. The answer agreed by the majority of pupils to each question was then inputted to the programme. This was repeated for each question. Teachers facilitated discussion around the questions and answers but did not influence the answers given by the pupils. The sessions were not generally delivered before science or maths lessons: only two of the 32 teachers responding to the survey said that this was the case in their school. Our observations of sessions revealed no set pattern across schools, for instance, one session was delivered in form time, another before a music lesson, and one was delivered after morning assembly before a spelling test. The training involved one face-to-face familiarisation session where the Birkbeck researcher demonstrated how to use the programme. The researcher was also available to deal with any follow-up queries from the teacher by telephone or email. Most of the teachers surveyed (26 out of 32 responding to the survey) said that the training was suitable in preparing them to use the See+ programme.

Evaluation objectives

This section describes the objectives of this evaluation. These are in line with the trial protocol (NFER, 2018) and the Statistical Analysis Plan (SAP; McNamara *et al.*, 2018).

Primary research question

- Does the use of the Stop and Think intervention impact on pupils' maths and science attainment?

We answered this primary research question by measuring pupil attainments in maths and science separately.

Secondary research question

- What is the effect of the Stop and Think intervention on pupil's inhibitory control function?

We answered this question by analysing the Chimeric Animal Stroop measure drawn from Wright *et al.* (2003).

Additional analysis compared the impact of the Stop and Think intervention on maths and science achievement to the impact of the social skills computer-based learning activity (See+) used by the control-plus group. This analysis helped to determine if any identifiable effect was due to using a computer programme (offered in a lesson other than maths or science) rather than any specific content. In addition to this, we also explored the effect of the See+ activity by comparing the primary outcome measures of the control-plus group and the business-as-usual control group (also referred as a control group).

The process evaluation aimed to investigate the following research questions:

- Was the theory of change model (Appendix E) identified in the pilot an accurate representation of the intervention and its outcomes?
- Have schools implemented the intervention in the way it was intended? If not, why not?
- Is the intervention appropriate for pupils of this key stage, of this age group, and in these lessons?
- Can programme materials and delivery be improved for the future?
- Is the roll-out of the intervention feasible for schools?

Ethics and trial registration

We obtained approval from NFER's Code of Practice Group on 16 March 2016. NFER used pupil administrative data from the Department for Education's (DfE) National Pupil Database (NPD). We matched the NPD data to the pupil assessment data collected by our test administrators and pupil personal data provided by the schools (via Birkbeck College).

Birkbeck College was responsible for school recruitment and the initial data collection. The headteacher (or a designated member of the senior leadership team) of the school made the decision whether to participate in the trial. They opted into the trial by signing a memorandum of understanding (MoU) during recruitment. Birkbeck College collected the name, job role, and contact details of the nominated staff member to liaise with for the purpose of this study. They also collected names and contact details of Year 3 and Year 5 teachers in order for NFER to conduct a survey and interviews during the evaluation. Prior to randomisation, schools also sent names, dates of birth, and Unique Pupil Numbers (UPNs) for participating Year 3 and Year 5 pupils to Birkbeck College. Schools sent parental opt-out consent letters prior to sending this data to Birkbeck College.³ The school information sheet, along with all of the other school communications, contained relevant information about consent and how the data was collected, matched, and stored.

Appendices C and D provide the school information sheet, the consent form, NFER's privacy notice, and the MoU.

³ Note that our legal basis for processing the personal data for this trial is our legitimate interest to administer the randomised controlled trial. See the section on data protection for further details.

The trial was registered with the ISRCTN registry as trial number: ISRCTN20284041. The registry is administered and published by BioMed Central.

Data protection

The General Data Protection Regulation (GDPR) became enforceable in May 2018. In March 2018, we shared the publicly available privacy notice (see NFER, n.d.) with all participating schools. This privacy notice included all relevant aspects of the personal data that we were collecting for this evaluation. The purpose of collecting the personal data for this trial was to ascertain the impact of the intervention on pupil attainment in maths and science. The legal basis for processing the personal data was covered by GDPR Article 6 (1f) which states that 'processing is necessary for the purpose of the legitimate interests'. Our legitimate interest for processing the personal data was to administer the randomised controlled trial as the evaluation fulfils one of NFER's core purposes (undertaking research, evaluation, and information activities). We did not collect any special data for this evaluation. Personal data is held by NFER and Birkbeck College. Both parties signed a data sharing agreement as joint data controller. The document also states that Birkbeck College will not have access to personal data that is provided by the NPD. As joint data controllers, both parties will delete any personal data three years after completion of the project.⁴ NFER will share all pupil data (pupil names, dates of birth, UPN matched to the NPD data described above, and assessment results) with the EEF's data archive manager, FFT Education, within three months of the end of the project. FFT Education will keep the data and take responsibility for data protection compliance.

Project team

The principal investigator for this trial was Simon Rutt, Head of NFER's Centre for Statistics. The day-to-day trial manager was Palak Roy, Senior Trials Manager (who took on this role in March 2016). Prior to this, the trial was managed by Dr Anneka Dawson (during her previous role at NFER). They were supported by Stephen McNamara, Sally Bradshaw and Afrah Dirie who undertook statistical analysis. The process evaluation was led by a team of researchers from NFER's Centre for Policy and Practice Research: Claire Easton and David Sims. The school communications were managed by researchers from NFER's Research Operations department: Tom Dickinson, Max Falinski, and Kathryn Hurd. The GL Assessment test administration was managed by Shalini Sharma and the tests were administered by trained NFER test administrators. NFER was responsible for the trial design (developed jointly with the delivery team) and for managing the ongoing relationship with the schools (jointly with the delivery team), as well as randomisation, analysis, and reporting of the independent evaluation.

The intervention was developed and delivered by a team at Birkbeck College and UCL-Institute of Education's Centre for Educational Neuroscience. It was led by Professor Denis Mareschal from Birkbeck assisted by Professor Michael Thomas, Dr Iroise Dumontheil, and Dr Hannah Wilkinson, and from UCL IoE by Professor Andie Tolmie, Professor Emily Farran, Dr Kasak Porayska-Pomsta, Dr Sveta Mayer. They were assisted by Professor Derek Bell from LEARNUS. Birkbeck College was responsible for school recruitment and administration of baseline and follow-up Stroop assessments.

The project was funded by the Education Endowment Foundation and the Wellcome Trust and was supported by EEF staff Dr Anneka Dawson (during her previous role at EEF), Camilla Nevill, Eleanor Stringer, and Dr Florentina Taylor.

⁴ Retention of personal data is subject to agreement by the NPD team at DfE.

Methods

Trial design

This was a three-arm cluster randomised controlled trial involving 89 primary schools. The three arms were *intervention*, *control-plus* (implementing a social skills programme called See+) and a business-as-usual *control group* that continued with usual classroom practice. The study included all Year 3 and Year 5 classes in participating schools. Randomisation was at the year group level and was stratified by form-entry of the schools in order to achieve balance across the groups. Within a school, Year 3 and Year 5 were randomly assigned to either the intervention group (to deliver Stop and Think at the start of maths or science lesson) or one of the two control groups. This randomisation process resulted in a ratio of 2:1:1 allocation to intervention, control, or control-plus groups. This meant that every school had intervention class/es for one year group and class/es from the other year group were randomly assigned to either the control group or the control-plus group. This ensured that each school received the intervention. Classes within each year group were always randomised to the same group. This design meant that there could be four possible scenarios of trial design. This is illustrated in Table 2.

Table 2: Trial design—four possible scenarios for group allocation

	Year group	Trial arm	Primary outcome tests
Scenario 1	Year 3	Intervention	PTM8
			PTS8
	Year 5	Control group	PTM10
			PTS10
Scenario 2	Year 3	Intervention	PTM8
			PTS8
	Year 5	Control plus group	PTM10
			PTS10
Scenario 3	Year 3	Control group	PTM8
			PTS8
	Year 5	Intervention	PTM10
			PTS10
Scenario 4	Year 3	Control plus group	PTM8
			PTS8
	Year 5	Intervention	PTM10
			PTS10

In Table 2, scenario 1, for example, illustrates the situation of Year 3 classes being randomised to the intervention group (Stop and Think) and Year 5 classes being randomised to the control group ('business as usual'). In order to make the outcome testing more efficient, pupils within each class were randomly allocated to take either maths or science test.

This way of within-school randomisation has two benefits: every school receives the intervention and fewer schools are required to be recruited to the trial. Primary schools tend to have a class teacher assigned to each class and therefore contamination was not an issue. Schools were not offered financial incentives to participate as each school had received the intervention. The trial ran according to the updated protocol published in 2018 (NFER, 2018). Table 3 presents the trial design in brief.

Table 3: Trial design

Trial type and number of arms		Three-arm cluster randomised controlled trial
Unit of randomisation		Randomisation at year group level
Stratification variable(s) (if applicable)		Number of forms—i.e. one-form, two-form, three-form and mixed-form entry schools
Primary outcome	variable	Pupil attainment in maths (combined across Year 3 and Year 5) Pupil attainment in science (combined across Year 3 and Year 5)
	measure (instrument, scale)	GL Assessment's Progress Test in Maths 8 (Year 3) GL Assessment's Progress Test in Maths 10 (Year 5) GL Assessment's Progress Test in Science 8 (Year 3) GL Assessment's Progress Test in Science 10 (Year 5)
Secondary outcome(s)	variable(s)	Pupil's inhibitory control function
	measure(s) (instrument, scale)	The Chimeric Animal Stroop measure drawn from Wright et al. (2003)

Participant selection

Birkbeck College was responsible for school recruitment. For the main trial, it recruited schools with predominantly, but not exclusively, an above average proportion of pupils eligible for FSM. The eligible schools also needed to have at least one Year 3 class and one Year 5 class. In other words, all primary schools were eligible to take part in the trial as long as they had at least two classes—one for each of these year groups. Schools with mixed year groups being taught in the same class were also eligible, so long as the Year 3 and Year 5 pupils were not being taught in the same class. For example, a school was eligible to take part in the trial if Year 3 pupils were being taught with either Year 2 or Year 4 pupils, but they were out of scope if the Year 3 pupils were being taught in the same class as Year 4 and Year 5 pupils. However, the trial only looked at Year 3 and Year 5 pupils which meant pupils from other year groups, despite being taught in the same class as Year 3 and/or Year 5, were not eligible trial participants.

In the five case studies we carried out as part of the process evaluation, we observed one mixed Year 3 and Year 4 class doing a Stop and Think session before a maths lesson. All of the pupils in the class participated in the Stop and Think session.

Schools were also required to provide administrative pupil data to Birkbeck College in order to be eligible for randomisation.

Birkbeck College used a multi-layered approach in order to recruit schools to take part in this trial.

- It sent a general email to schools using a national database. It contacted a large number of schools but the response rate was very low.
- It approached schools with which it had personal contacts. Although the number of schools reached this way was relatively smaller, it yielded a good response rate.
- It used its own contacts with individuals in targeted areas of the country—namely Wirral, Manchester, Birmingham, Sheffield, and South West England (Devon, Cornwall, Somerset, and Bristol). It also reached out to schools via local newsletters and existing maths and science networks. On the whole, this produced a good response rate, although the strength of response rate fluctuated from one area to another.
- It approached schools via national networks and newsletters including Primary Science Teaching Trust, Primary Science Quality Mark, and the National Education Trust. It was quite difficult to monitor the response rate of this approach. Although the direct response rate appeared to be low, this method produced some strong contacts who helped to engage some local networks of schools.

Besides this, Birkbeck College also sent targeted emails to schools, in particular towns and areas, which yielded relatively higher response rates than untargeted mass mailing. In addition, it used some social media networks; this yielded relatively low responses. While there are some schools that will have received more than one email from

Birkbeck College, it is estimated that over 10,000 schools were contacted, either via email or through personal/network contacts. These communications generated an expression of interest from 250 schools.

Outcome measures

Primary outcome

There were two primary outcome measures for this trial:⁵ pupil attainment in maths and pupil attainment in science. After revising the original protocol, it was decided to retain both the primary outcome measures. However, the use of two primary outcomes increases the risk that a false positive result may be found through chance. And in order for the trial to demonstrate an effect, 95% confidence intervals from separate maths and science analyses must not overlap with zero (EEF, 2018, p.6).

The primary outcomes were measured by administering the Progress Test in Maths (PTM) and the Progress Test in Science (PTS) produced by GL Assessment. PTM is a standardised assessment of pupils' mathematical skills and knowledge including number, shape, data handling, and algebra. PTS is a standardised assessment of pupils' science knowledge including the three core areas of physics, chemistry and biology as well as 'working scientifically'. Further details on PTM and PTS links to national curricula and technical details of the assessments can be found from GL Assessment website.⁶

NFER managed the test administration by sending independent test administrators into schools that were taking part in February and March 2018. This helped to ensure that the tests were administered blind to randomisation group allocation⁷ and reduced the burden placed on schools by ensuring that the teachers did not have to administer the tests. As there were two year groups taking part in the trial, it was necessary to administer age-appropriate tests. Year 3 pupils took PTM8 and PTS8 and Year 5 pupils took PTM10 and PTS10. We organised for more than one test administrator to visit each school in order to accommodate the testing of two year groups taking two separate assessments. The power calculations were based on each pupil taking only one subject test. An NFER statistician assigned subject tests to pupils using a simple randomisation such that equal numbers of pupils within a class were randomly allocated to take a maths or science test. This list of pupil test allocations was sent to schools one week prior to testing to facilitate efficient test administration.

We used raw total scores from the PTM and PTS tests as the primary outcome measures. The maximum possible score is 55 for PTM8, 65 for PTM10, 40 for PTS8, and 50 for PTS10. On all of these assessments, a higher score indicates higher attainment. As Year 3 and Year 5 pupils took different assessments, it was necessary to analyse outcomes from these assessments separately. For example, for maths, outcomes from PTM8 and PTM10 were analysed using separate models. In order to determine the impact of the intervention on pupil attainment in maths, we needed to combine maths attainment outcomes from both the year groups. This meant that the effect sizes from PTM8 and PTM10 models were combined to determine an overall impact of the intervention on pupils' attainment in maths. This combined effect size constituted the primary outcome measure in maths. Similarly, we combined effect sizes from PTS8 and PTS10 to constitute an effect size for the primary outcome measure in science. Further details on how we combined the effect sizes can be found in the section on 'Statistical Analysis'.

Secondary outcome

The secondary outcome measure for the trial was assessed using the Chimeric Animal Stroop measure of inhibitory control. This assessment was chosen by the delivery team and was adapted from Wright et al. (2003). It was a pencil-and-paper version which allowed a whole-class assessment in schools that did not have individual child computer facilities available. All children carried out the same pencil-and-paper version for consistency. Pupils worked through five sheets: one practice, two congruent conditions, and two mixed conditions. Pilot testing with a group of ten primary

⁵ Note that there were two primary outcome options in the original protocol. After the development phase was complete, it was decided that both the outcomes would be retained as the primary outcome measures for the trial as suggested in the original protocol.

⁶ <https://www.gl-assessment.co.uk/support/ptm-product-support/> and <https://www.gl-assessment.co.uk/support/pts-product-support/>

⁷ Although the test administrators were from NFER, they did not have access to the randomisation results and therefore were blind to group allocation.

school children (aged 5–11 years) was undertaken to test the instructions were clear and to set a time limit for the task. As a result, ten seconds per sheet was set to avoid floor or ceiling effects in either block across year groups. The same task was used for both year groups.

Pupils were asked to recognise an animal's body in a picture while ignoring the head and needed to complete as much of each sheet as possible within ten seconds. In the congruent condition, the animal head matched with the animal body. In the mixed condition, half the animals had heads that matched their bodies and the other half had heads that did not match their bodies. It was the latter that enabled us to assess their absolute performance in the mixed conditions. The raw total score from the mixed sheets was used as a secondary outcome measure for the analysis. This score ranged from 0–30, where a higher score indicated better attainment. The raw total score from the congruent sheets was included in the model to control for cognitive skills not related to inhibitory control. This test was administered by the research assistants appointed by Birkbeck College in March–April 2018. This was after the schools had completed the GL Assessment tests. The Birkbeck research assistants who carried out the Stroop assessments were allocated to schools that they had not previously worked with so they were blind to the class's intervention condition.

Sample size

Sample size from the protocol

Initially, the required number of schools was not driven by the sample size calculations. Prior to NFER's involvement, the EEF's project summary presented 100 schools being required for the trial. It was acknowledged that a design based on 100 schools would result in a relatively low Minimum Detectable Effect Size (MDES). Sample size and MDES calculation at the time of writing the protocol, did not consider separate models for Year 3 and Year 5. MDESs were calculated for each subject by dividing the available pupil numbers in half. We did not account for multiple testing as both the subject outcomes were retained as the primary outcome measures. This is considered a conservative approach and setting the bar too high, given that in order for the trial to demonstrate an effect, we needed to reach the statistical significance in both the subject outcomes.

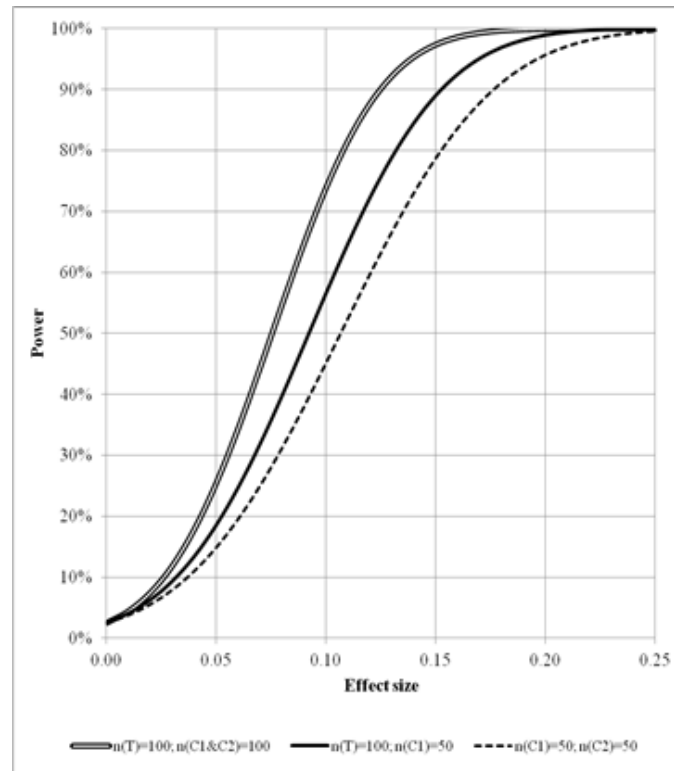
At protocol, we used a power calculation with the following two assumptions obtained from the EEF's paper on pre-test effects (EEF, 2013): correlation between pre-test and post-test would be 0.65 and the intra-class correlation would be 0.126. The initial design (in the protocol) presented Key Stage 1 (KS1) attainment measures to be used as a covariate. But due to the changes in how KS1 is measured and reported, we could no longer use KS1 as it would not be consistent for both year groups. Instead, in discussion with the EEF, we decided to use the Early Years Foundation Stage Profile (EYFSP) as a pre-test measure for both primary outcome measures—maths and science. We presented this change in the SAP (McNamara, 2018).

The correlation between EYFSP and the primary outcome measures in Year 3 and Year 5 were assumed to be 0.65. The rationale for selecting a correlation of this size was based on a paper produced by the Fisher Family Trust (FFT, n.d.). These figures were used in the calculation of optimum sample sizes for desired levels of power. These assumptions allowed for the following comparisons:

- *Primary outcomes in each subject:* n (intervention) = 100 schools and 150 classes; n (control and control-plus) = 100 schools and 150 classes represented the comparison between intervention classes and both control and control-plus classes grouped together and assumed an average cluster size of 27 (average cohort size for eligible primary schools class in England). Power calculations were based on half of these pupils taking a maths test and the other half taking a science test as pupils were going to be randomly allocated to either subject tests. Calculations were based on an effect size for either of these tests. Both assessments were therefore powered to 80%. The power curve is demonstrated in Figure 1 as a comparison of $n(T)$ and $n(C1\&c2)$ with a minimum detectable effect size (MDES) of 0.125.
- *Comparison of intervention vs. control-plus group:* n (intervention) = 100 schools and 150 classes; n (control-plus) = 50 schools and 75 classes represents the comparison between the intervention classes and the control-plus group. This again assumes an average cluster size of 27 (average cohort size for eligible primary schools in England). Each of the two subjects were powered to 80%. This is demonstrated in Figure 1 as a comparison of $n(T)$ and $n(C1)$.

- *Comparison of control-plus vs. control group:* n (control) = 50 schools and 75 classes; n (control-plus) = 50 schools and 75 classes represents the comparison between the control and control-plus groups. This assumes an average cluster size of 27 for the size of each class. This is demonstrated in Figure 1 as a comparison of $n(C1)$ and $n(C2)$.

Figure 1: Power curves for three comparisons, from the protocol



While writing the protocol, it was assumed that there were 22.5% of pupils who were eligible for FSM at any time during the previous six Years ($n=4.56$). Therefore, the MDES for FSM only analysis was estimated to be 0.17 at 80% power.

MDES at randomisation

The evaluation team randomised 89 schools. Following the recruitment and randomisation, two schools withdrew from the trial without the knowledge of group allocation. One more school withdrew from the primary outcome testing. Therefore, at the time of writing the SAP the total number of schools in the final analysis were assumed to be 86. We revised the sample size calculation based on the number of schools and number of pupils available for primary outcomes testing. The MDES for each subject analysis was 0.135 with assumptions similar to those discussed above (please see Table 4 for further details on number of pupils and schools). We did not have access to pupil FSM eligibility at the time of writing the SAP. Assuming that 22.5% of the cohort would be FSM eligible ($n = 4.18$), revised MDES for FSM-only analysis was increased slightly to 0.19 with 80% power. MDES at various stages is presented in Table 10.

Randomisation

Birkbeck College recruited 97 primary schools. Schools signed up to the trial by signing the MoU (Appendix D). Of these, five schools had only one Year 3 or Year 5 year group or a mixed class across both year groups. Therefore, these schools were not eligible to take part in the trial. Additionally, two schools did not submit their administrative pupil data and one school withdrew participation prior to randomisation. As a result, we randomised 89 schools (178 year groups).

An NFER statistician carried out the randomisation using SPSS with a full syntax trail. The syntax is included in Appendix F. Two waves of randomisation took place during October 2017 to facilitate the training. Two waves were

necessary to ensure that Birkbeck College could start with the software installation and training for Wave 1 schools while the rest of the schools sent their administrative data. Research assistants from Birkbeck College delivered the training, which included a visit to all the participating schools and installing the computer programme(s). The software installation took place in all schools as at least one class in each school would be allocated to the intervention. There was no need for baseline testing as we used the EYFSP as a prior attainment measure in the analysis. EYFSP is created based on pupils' assessment at the end of their early years (when they are four and five), which means the baseline data was collected prior to randomisation.

The randomisation was stratified by the number of form entry (that is, number of classes in each year) as it was important to ensure that the number of intervention classes was similar to the number of classes in the control group and the control-plus group together.

The trial schools had a variety of form-entry structures and these were different from the one- and two-form-entry scenarios originally specified in the protocol. This structure needed to be accounted for in the randomisation process. Schools were randomised as one of four possible set-ups:

- one-form entry, both years;
- two-form entry, both years;
- three-form entry, both years; and
- all other form entry (that is, one Year 3 class and two Year 5 classes, two Year 3 classes and three Year 5 classes, four Year 3 classes and four Year 5 classes).

Table 4 presents the number of schools and year groups randomised to each trial arm. As explained earlier, for every school, we assigned one Year group to the intervention group and another Year group to either of the two control groups. For example, for the first wave, in 28 schools, Year 3 was randomised to the intervention group, in 15 schools Year 3 was randomised to the control group and in 16 schools Year 3 was randomised to the control-plus group. As Primary schools tend to have a class teacher assigned to each class, contamination was not an issue. The rationale for choosing the Year group as a unit of randomisation is included in the *Trial Design* section above.

Overall, there is an imbalance in the group allocation for Year groups in the control and control-plus groups. This occurred as a result of not correcting the group imbalance that arose in the first wave. We deliberately adopted this approach to ensure that bias is not introduced in case the schools in the second wave are systematically different from the schools in the first wave. By not correcting this imbalance, we are not allowing more 'control groups' than 'control-plus groups' in the second wave schools.

Shortly after randomisation, two schools withdrew participation from the trial. These schools did not know their group allocation and therefore we removed them from the trial and subsequent analysis. The final number of schools retained in the trial was 87.

Table 4: Number and proportion of schools and year groups at randomisation

	Trial arms	Year 3	Year 5	Total Year Groups
Wave 1: 59 Schools, 118 year groups	Intervention	28	31	59 (50%)
	Control	15	13	28 (24%)
	Control Plus	16	15	31 (26%)
Wave 2: 30 Schools, 60 year groups)	Intervention	14	16	30 (50%)
	Control	7	7	14 (23%)
	Control Plus	9	7	16 (27%)
Total: 89 Schools, 178 year groups	Intervention	42	47	89 (50%)
	Control	22	20	42 (24%)
	Control Plus	25	22	47 (27%)

Statistical analysis

The analysis followed the EEF's Statistical Analysis Guidance (EEF, 2018) and the trial SAP (McNamara, 2018). This section provides an outline of the analysis undertaken.

Primary intention-to-treat (ITT) analysis

We ran a separate analysis for each primary outcome measure—one for maths and one for science. The overall impact of the intervention on pupils' attainment in a given subject was determined by combining the effect sizes from the two year group models. For maths, we analysed outcomes from PTM8 (Year 3) and PTM10 (Year 5) in two separate models. The combined effect size constituted the primary outcome measure in maths (please see the section on effect size calculation). We also ran similar models for science and combined them to determine an overall effect size for the primary outcome measure in science. Model details and the calculation of effect sizes are described below:

The primary outcomes' analyses were 'intention-to-treat', and were conducted at pupil level, comparing an average pupil maths or science score in the intervention group with an average score in the combined control (control and control-plus) group. As the pupil-level data was clustered within classes, which were clustered within Year groups and schools, the hierarchy of the data needed to be acknowledged in the models. We ran each model at year-group level so year group was not included as one of the levels. Hence, multilevel linear regression models with three levels (school, classes, and pupils) were used to analyse the impact of the intervention on pupil outcomes.

As acknowledged in the SAP (McNamara, 2018), we could not use KS1 assessment data as pupil prior attainment due to the changes in the data collection and reporting arrangements in the NPD. Instead, we used average EYFSP point scores by combining all 17 early learning goals. These variables were available on the NPD with a value range of 1–3 where higher scores reflected higher attainment for a given goal. Further details on the distributions of the prior attainment measures can be found in the analysis section.

Maths primary outcome

As mentioned previously, we ran two separate models for maths—one for each Year group. In Year 3 maths model, the dependent variable was the PTM8 raw total score with the following covariates:

- an indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups);
- the stratification variable used at randomisation to indicate whether the school is a two-form-entry, three-form-entry, or mixed-form-entry school (reference category = one-form-entry school); and

- the pupil's average EYFSP score as a prior attainment measure.

The Year 5 maths model also followed similar structure where the dependent variable PTM10 raw total score was regressed on the same covariates.

Effect size calculation

The numerator for each individual model effect size calculation was the coefficient of the intervention group from the multilevel model. These effect sizes were calculated using the total variance from the multilevel models without covariates as the denominator, that is, equivalent to Hedges' *g*. Confidence intervals for each effect size were derived by multiplying the standard error of the intervention group model coefficient by 1.96. These were converted to effect size confidence intervals using the same formula as the effect size itself.

The overall effect for the maths outcome was an amalgamation of the effects of Year 3 and Year 5 models. These were different pupils and the only non-independent component in the analysis was the school effect which was already taken into consideration while running the multilevel models. Therefore, we combined the effect sizes from these models according to the method described by Borenstein *et al.* (2009, p. 218) and applying formulas from 11.3 and 11.4 from page 66 of the same work. This method allowed the combination of effects from independent subgroups.

Here are the formulas used to amalgamate the two effects sizes where, for the Year 3 and Year 5 models respectively, Y_{m3} and Y_{m5} were the effects sizes, V_{m3} and V_{m5} were the variances, and Y_{cm} and V_{cm} were combined effect size.

Firstly, we calculated the weights assigned to each model.

$$W_{m3} = \frac{1}{V_{m3}} \text{ and } W_{m5} = \frac{1}{V_{m5}}$$

Where, W_{m3} and W_{m5} were the weights for Year 3 and Year 5 maths models respectively.

Combined effect size Y_{cm} was:

$$Y_{cm} = \frac{(Y_{m3} * W_{m3}) + (Y_{m5} * W_{m5})}{W_{m3} + W_{m5}}$$

Combined variance V_{cm} was:

$$V_{cm} = \frac{1}{W_{m3} + W_{m5}}$$

Science primary outcome

Similar to maths, we ran two models for the science outcomes—one for Year 3 and one for Year 5. The models included the same set of covariates as described above. We calculated the combined effect size and the combined variance following the same methods described above.

Imbalance at baseline for analysed groups

Imbalance in the group allocation (as assigned at randomisation) was explored with regards to background characteristics such as pupil FSM eligibility and prior attainment. We used multilevel models with three levels (schools, classes, and pupils) to examine the imbalance at baseline using EYFSP as prior attainment. We ran four separate models, one for each year group and subject. Prior attainment was regressed on whether the pupil belonged to the intervention or the combined control groups as well as including the stratification variable used at randomisation. The

presence of imbalance was determined by calculating the effect sizes for each of the four models. The findings from this analysis are presented in the impact evaluation section and Table 12.

Missing data

As per the SAP, we assessed missing data at the randomisation level of year groups. As seen in the participant flow diagram (Figure 2, page 30), 87 intervention year groups were meant to be followed up; of these, 84 took part in primary outcome testing. This meant the attrition for the intervention group was 3%. Eighty-seven year groups were meant to be followed up from the combined control groups and, similar to the intervention group, we had outcomes data from 84 year groups and therefore the attrition from the combined control group was also 3%. We also explored the attrition at pupil level as the unit of analysis was pupils. Pupil-level attrition is presented in Table 9. We ran four multilevel logistic models (one for each GL Assessment outcome) with three levels (school, pupil, and classes) on whether or not a pupil was missing at follow-up, regressed on a number of covariates in addition to the ones in the main model. Since the pupil-level attrition was higher than 5%, it was important to explore the level of missing data and the extent of bias. To do so, we ran multilevel multiple imputation and compared results from the complete data analysis with the ITT models.

Complier Average Causal Effect (CACE)

As the intervention was computer-based, it was possible for the delivery team to extract an exact number of sessions completed by each class. The main analysis was, therefore, followed by a CACE analysis in order to assess the effect of non-compliance on outcome measures where the data from the computer system was used to determine the extent of each class's involvement. The delivery team, as agreed with NFER, collected the information on a number of completed sessions for each class. This determined the compliance or engagement level of each class. This data assumed that pupils from a given class would have received the same number of Stop and Think sessions, which enabled us to determine the number of Stop and Think sessions each pupil participated in and whether it constituted compliance at pupil level. In doing so, the compliance measure did not take pupil absence into consideration. In the SAP, we envisaged that we would use ordered categories to describe low, medium, and high level of compliance. However, we used the number of sessions as a continuous measure of compliance as this variable provided us with more information.

We used a two-stage least-squares model to calculate the CACE estimate (Angrist and Imbens, 1995). We ran four separate models, one for each year group and subject. In the first stage of the model, we regressed pupil-level compliance (as above) on all covariates used in the main primary outcome model and included (as an instrumental variable) a binary variable that indicated a pupil's pre-intervention treatment allocation. The second stage of the model regressed the primary outcome on the covariates used in the main models and included a covariate representing pupil's estimated level of compliance, which was derived from the first stage of the model. The coefficient of the compliance measure was the CACE estimate. We used the R package ivpack to perform the CACE analysis on the primary outcomes only.

Secondary outcome analysis

Four multilevel models with GL Assessment scores as outcome measures constituted the secondary analyses. These models are described in the primary analysis section, as they were required to calculate the combined effect size that constituted the primary outcome measure for each subject.

The outcome of the Chimeric Animal Stroop task was analysed via the multilevel linear regression model. We performed analyses at pupil level, in a three-level hierarchy to account for the clustering within classes and schools. There were two separate models—one for each year group. As discussed in the section on secondary outcomes, the dependent variable in these models were the raw total score from the mixed conditions sheets from the Stroop task. This was regressed on the following covariates:

- an indicator of whether the pupil was in the intervention group (reference category = combined control group that consists of both control groups);
- the stratification variable used at randomisation to indicate whether the school is a two-form-entry, three-form-entry or mixed-form-entry school (reference category = one-form entry school); and
- the raw total score in the congruent sheets as a control for non-inhibitory control cognitive skills.

For further details on the congruent conditions and mixed conditions, see the section on secondary outcome. The combined effect size from the two year-group models determined the overall impact of the intervention on this outcome of inhibitory control. As discussed earlier, these were combined according to the method described in Borenstein *et al.* (2009).

Additional analyses

We undertook two additional analyses. For each analysis, we ran separate models for each year group and subject using data from a subset of pupils in two of the three arms of the trial. The first analysis looked at the differences between the intervention and the control-plus group. Outcomes from this analysis helped us to determine whether—if an impact of the intervention was seen—it was purely due to introducing a novel computer programme or could be specifically attributed to the Stop and Think intervention. The second analysis looked at differences between the control-plus group and the business-as-usual control group. The model structures were similar to those discussed in the primary analysis.

Subgroup analyses

As specified in the SAP, subgroup analyses took place to explore the differential impact of the intervention when pupils' FSM and gender were taken into consideration.⁸ This was done using interaction models that were identical to the primary outcomes models but including the variable of interest (everFSM or gender) and the variable interacted with the intervention as additional covariates. Analyses then proceeded as per the original primary outcomes models.

We also ran separate analyses of everFSM pupils as per EEF requirement. We used the EVERFSM_6_P_SPR18 variable from the 2017/2018 Spring school census. These models were identical to the primary analyses models except that they only included everFSM pupils.

All the data manipulation took place in IBM SPSS Statistics 24 and the multilevel models were run in R version 3.3.3 and above.

Implementation and process evaluation

The purpose of the implementation and process evaluation was to provide information on the programme and insights into its delivery; it evaluated the pilot (Phase 1) and the main trial (Phase 2). Phase 1 involved collecting information on:

- the models of delivering the intervention being used;
- the feasibility of delivering the intervention;
- the teacher training related to the intervention;
- any other maths/science interventions taking place in the schools; and
- any other neuroscience-based interventions taking place in the schools.

In Phase 1 we conducted visits to two schools to interview senior leaders and teachers involved in the implementation of the UnLocke maths and science programme. In the first school we interviewed the deputy headteacher and a Year 3 teacher; in the second school we interviewed the headteacher and a Year 5 teacher. We also received email responses to our schedule of questions from a senior leader in another school which was delivering the See+ programme.

Phase 2 investigated the following research questions:

- Was the TOC model identified in the pilot an accurate representation of the intervention and its outcomes?
- Have schools implemented the intervention in the way it was intended? If not, why not?
- Is the intervention appropriate for pupils at this age and in these lessons?

⁸ The evidence about gender difference in maths and science attainment is not consistent. A decision was made at the protocol stage that we will analyse whether the effectiveness of the intervention is differential by gender.

- Can the programme materials and delivery be improved in the future?
- Is the roll out of the intervention feasible for schools?

In Phase 2 we carried out visits to five schools and interviewed four senior leaders and four teachers and undertook observations of Stop and Think sessions in four schools and observations of See+ sessions in two schools. We conducted post-intervention telephone interviews with teachers from six trial schools (one interview was conducted with the control-plus teacher, not the intervention teacher).

A summary of the Phase 1 and Phase 2 fieldwork interviews is presented in Table 5 below.

Table 5: Summary of Phase 1 and Phase 2 implementation and process research interviews

Phase and research activity	Number of schools Stop and Think Year 3	Number of schools Stop and Think Year 5	Teachers interviewed	Senior leaders interviewed
Phase 1 Pilot: 2 schools, face-to-face interviews	1	1	2	2
Phase 2 Case-study visits to 5 schools, face-to-face interviews	3	2	4	4
Phase 2 Post-intervention telephone interviews, 6 schools	2	4	5	1

Researchers selected the research methods outlined above as they offered both breadth and depth to the implementation and process evaluation. We considered that the methods were appropriate for this trial because they enabled us to examine how schools delivered a computerised learning programme for Stop and Think and See+ in the school setting. Through using these methods we gained interview and observational evidence of how well the delivery worked and what delivery challenges schools encountered and how they addressed them. This yielded valuable data on the feasibility of the roll-out of the intervention.

The implementation and process evaluation offered further insight into the impact of Stop and Think on pupils' maths and science achievement and on pupils' inhibitory control by providing teachers' perceptions of impact.

Findings from phase 1 Pilot interviews

Stop and Think and See+ software programmes were developed and piloted during the development phase by Birkbeck College. It selected eight primary schools to take part during this phase. Subsequently, three schools dropped out of the pilot delivery: one school faced staff turnover problems and the other two schools reported having difficulties with their school's IT systems.⁹ The remaining five pilot schools delivered the Stop and Think programme in two different ways: three schools adopted the whole-class approach—the teacher delivered the programme to the whole class—and the remaining schools adopted the individual approach where each pupil accessed the programme using a computer to work through the programme individually.

As highlighted before, the recruitment for the main trial commenced before the pilot phase was completed. Therefore, the decision about the preferred mode of delivery was based on the developer's own experience of implementation rather than NFER's process evaluation.¹⁰ The view was that providing individual computers or laptops to all class pupils was a substantial logistical challenge for the majority of schools and could inadvertently affect the participation rates for the main trial. Therefore, it was decided that the main trial would include delivering the intervention and control-plus software in a whole-class setting.

⁹ Birkbeck College investigated the IT issues further. These were to do with school's own systems rather than accessing the software.

¹⁰ However, the process evaluation findings were broadly in line with the developer's findings, where relevant.

In March 2017, we visited two schools to interview teachers and senior leaders involved in a pilot implementation of the Stop and Think programme.¹¹ The schools adopted two different approaches:

- One school used the programme with Year 3 in a whole-class format, facilitated by the teacher from the front of the class using an interactive whiteboard. This school was able to adhere to three 15-minute sessions per week at the start of a maths or science lesson. Mostly, Stop and Think was used at the start of maths lessons.
- The other school introduced the programme to all Year 5 pupils with each pupil using a laptop to work through the programme individually. While the school managed to deliver three sessions per week, these tended to last around 40 minutes (due to the time commitment to set up the laptops). On the whole, UnLocke was not used in maths or science in this school and instead was delivered in the afternoon outside of core learning lessons.

The main findings on the accessibility and appropriateness of the Stop and Think programme were as follows:

- Interviewees said the Stop and Think programme is accessible and appropriate for all pupils including those with learning difficulties. In school one (whole-class approach) it particularly helped less confident pupils through the concept of 'stop and think' and the repetitious format. It was perceived to have additional impact due to the cooperation between the children, discussions, and language development.
- In both schools, pupils using the programme understood the content and format, especially the 'stop and think' message. 'Stop and think' was particularly successful in one school and had been introduced to other parts of school life (for example, other lessons and in the school council). Both teachers felt the 'stop and think' concept potentially had wider benefits (for example, enhanced resilience and application for life outside of lessons).
- While schools agreed that the content was aligned to the curriculum, the sequencing of questions did not align with the sequence of topics taught in school. For one school, this caused a challenge, in part resulting in them using UnLocke to reinforce learning rather than using it as part of a lesson. Within the other school (whole-class approach), the teacher used UnLocke instead of the usual warm up. While some of the UnLocke science content was new and unfamiliar to pupils, it helped this school introduce science concepts and language more regularly within the school week. Where the children had already learned a topic, they were excited to be able to answer questions but reportedly were not discouraged by new or non-familiar topics.

Both schools experienced technical challenges, however, the extent to which this caused problems was determined by the approach adopted. For the school which adopted the whole-class approach, the screen froze twice but this did not cause difficulty as the teacher was able to continue the lesson until the system unfroze. The other school needed to purchase new hardware, update its Wifi, and spend a significant amount of time installing the software on multiple laptops. In addition, it experienced regular screen freezes due to multiple laptops accessing the internet at once. Within both schools, the IT technicians had raised initial concerns about installing software onto the school's hardware but this was quickly overcome.

Interviewees in both schools were positive about Birkbeck's support. Both felt the training was sufficient and that the programme could be learnt on the job. One teacher noted the benefit of the informal training during a break time which meant that teaching time was not lost, neither was supply cover needed—a logistical challenge for schools.

In September 2017, we contacted three schools to receive feedback on See+ and received written feedback from a senior leader in one of these schools. All participants were assured anonymity. The feedback on See+ indicated that the school experienced some technical difficulties with using the programme which limited the progress of each session and the programme was considered too basic for Year 5 pupils.

The purpose of the development phase of the intervention and pilot research study was to explore the feasibility and scalability of the programme. Furthermore, researchers sought to gather feedback on whether having two groups (intervention and control-plus or control classes) within one school was practical. We found this was not an issue as the three conditions were in different Year groups. The findings from NFER's pilot activity were shared with the EEF, the Wellcome Trust, and Birkbeck College.

¹¹ We contacted two further schools to invite their participation in the research but, unfortunately, we were unable to secure visits or telephone interviews with participating teachers/senior leaders.

Phase 2 case studies

In October 2017, we selected a sample of six schools (five delivering Stop and Think and one delivering See+), ensuring a range of region/trainer (each region has a dedicated Birkbeck College trainer) and Ofsted ratings were covered. We drew the sample, selecting the first school in the list based on region/trainer and then Ofsted rating to ensure we included schools from across the Ofsted categories. Two of the originally selected schools were unable to participate as case studies and we found replacement schools that matched the intervention year groups and region/trainer. We had difficulty securing a visit to the sixth case-study school despite approaching a further four similar schools. Reasons for non-participation related to timing (term times, and the intervention having been completed or delayed in starting).

Between December 2017 and March 2018, NFER researchers visited five schools and interviewed four class teachers (three Year 3 teachers and one Year 5 teacher) and four senior leaders. The teachers participating in the interviews were those who had direct experience of delivering Stop and Think or See+ and could therefore give feedback on the practicalities of running the programmes in their school and class context. We also carried out four observations of the Stop and Think intervention and three observations of See+ (control-plus). All participants were assured anonymity. In the observations we looked at the delivery of the session including how the teacher introduced the session, whether it was a whole-class activity, how the teacher facilitated it, how pupils reacted, and whether any practical problems were encountered.

We analysed the case-study data treating the school as a whole unit. This involved examining the strategic account of how participating in the intervention fitted in with the school's educational priorities and the operational account given by the teacher focusing on practical issues of delivery. We also examined how the teacher's account compared with what we saw in our observation of a live session. Then we looked across the case-studies to identify common issues and any differences in approach and experience.

Table 6: Observations of intervention and control-plus

	Year 3 classes	Year 5 classes
Stop and Think	one mixed Year 3/4 class two Year 3 classes	one class
See+	one mixed Year 3/4 class	two classes (although the software did not run in one observation)
Total number of observations across five schools	4	3

The case-study interviews with teachers and senior leaders were semi-structured and lasted 20–30 minutes. As an introduction, we explained the purpose of the project, identified the topics we were going to cover, and guaranteed confidentiality. We also asked whether the interviewee would give permission for the interview to be audio-recorded. Interviewees were asked at the end of the interview if they had anything further to add.

The interview topics covered:

- why schools got involved in the Unlocked Project;
- how the project had been implemented within their school;
- views on feasibility of running a programme of this nature (three times a week for ten weeks);
- views on the training and support provided by Birkbeck researchers;
- perceptions of impact on pupils and on teachers;
- what worked well in terms of delivering the programme;
- barriers, challenges, and suggestions for improvement;
- the time and cost associated with delivering the programme in school;
- whether the programmes were age appropriate and the content was suitable for all primary-phase year groups;
- the extent to which pupils and teachers talked about the intervention and referred to the concept of pausing before answering in other situations within school; and

- recommendations for roll-out.

End-of-intervention survey of teachers

The purpose of the online survey was to gain a broad overview of the implementation of Stop and Think and See+ including any barriers teachers may have experienced and any outcomes observed by teachers on pupils and their teaching. The combination of the teacher survey and the case studies was designed to provide a full understanding of how and why the intervention has, or has not, worked including implementation challenges and adaptations, any unexpected outcomes, perceived impacts and benefits, and teachers' views on its sustainability and suitability for national roll-out.

In March 2018, NFER administered an online survey to all 87 schools involved in the trial. We asked the participating school contact to forward the survey to all teachers involved in either Stop and Think or See+ in their school. There was one online survey which included a set of questions for each programme. The respondents were routed to questions specific to the programme that they facilitated. We received responses from 63 schools and 105 teachers altogether. Of these, 61 teachers facilitated Stop and Think session, 32 See+ session, and 12 respondents had not delivered either programme and were asked no further questions. We issued two reminder emails and telephone reminders. We assured anonymity to all respondents. The teacher survey is presented in Appendix G.

Table 7: Response rates to teacher surveys

	Unlocked Project Schools	Teachers delivering Stop and Think	Teachers delivery See+
Number of completed questionnaires	63	61	32
Response rate (school n = 87)	72%		

Post-intervention telephone interviews

We carried out post-intervention telephone interviews with teachers. The purpose of these interviews was to gain additional feedback on what teachers thought about the Stop and Think or See+ programmes. During April 2018, researchers drew a random sample of 20 schools across the country that had not been involved in the case studies though they might have participated in the online survey and invited them to participate in a short telephone interview about their experiences of the programme. Researchers conducted six interviews with teachers, one in each of six schools, and carried these out during May 2018. Three of the schools delivered Stop and Think to Year 5 pupils, two schools delivered Stop and Think to Year 3 pupils, and one interviewee wanted to talk about delivery of See+ to the school's Year 3 pupils.

The interviews covered similar issues as explored with the case-study schools, focusing on:

- how the programme was implemented in school and whether it was suitable for all year groups;
- how the programme aligned with the maths and science curricula and what other maths and science interventions were delivered;
- perceptions of impact, what worked well and what could be changed; and
- the time associated with running the programme in school.

All participants were assured anonymity.

In our overarching analysis of interview data, observation data, and survey data we examined the responses to common topics in the schedules and questionnaire. We identified the numeric values from the survey for each topic and cross-referenced the responses with the interview and observation data to present a holistic quantitative and qualitative picture of intervention delivery. This approach enabled us to provide a broad and in-depth evidence-based account and assessment of the delivery of the intervention, drawing out messages for future roll-out.

Costs

We reviewed the delivery cost with the team at Birkbeck College. These are summarised in the findings sections. Birkbeck College developed a beta version of the Stop and Think computer programme during the development phase. This version was used for the evaluation.

We also examined training and implementation costs that were borne by schools. Some of these were not applicable to the intervention as the mode of delivery was online and teachers were to act only as a ‘facilitator’. The training and implementation took place in the school, which meant there were no travel or subsistence costs. As some pilot schools had reported that they had needed to purchase new hardware, update their Wifi, and spend a significant amount of time installing the software on multiple computers, we explored whether this was the case in the main trial. We asked this directly in the teacher survey. As the intervention was online and teachers did not require printing or photocopying material, there were no other additional costs borne by schools.

The teacher survey also explored the time teachers spent undertaking activities related to preparing for and delivering the Stop and Think sessions.

Timeline

The evaluation was set up in the autumn term of 2015. It is divided into two phases: (1) 18-month development and pilot phase and (2) a randomised controlled trial phase (main trial). The software programmes were developed and piloted during the development phase. NFER undertook a small-scale pilot process evaluation during this time. However, recruitment for the main trial commenced prior to completion of the pilot phase.

Table 8: Timeline

Date	Activity
Sep–Dec 2015	Meeting with partner organisations Draft the trial protocol
Jan 2016–Jul 2017	Development and pilot phase Software development by Birkbeck College Pilot the intervention—Birkbeck College Process evaluation interviews with the pilot schools—NFER
Jan 2017–Jul 2017	School recruitment for the main trial—Birkbeck College Collect relevant school and pupil administrative data (main trial)—Birkbeck College
Sep–Oct 2017	Update pupil administrative data—Birkbeck College
Oct 2017	Randomisation of schools—NFER Installation of computer programmes (software/s) in schools—Birkbeck College Training of teachers—Birkbeck College
Nov 2017–Feb 2018	Implementation of Stop and Think and See+ programmes Process evaluation case studies—NFER
Feb–March 2018	Process evaluation phone interviews—NFER Administration of teacher surveys—NFER Primary outcomes test administration (Progress Test in Maths and Progress Test in Science)—NFER
March–April 2018	Secondary outcome test administration (The Chimeric Animal Stroop)—Birkbeck College
Oct 2018–Feb 2019	Analysis and reporting—NFER

Impact evaluation

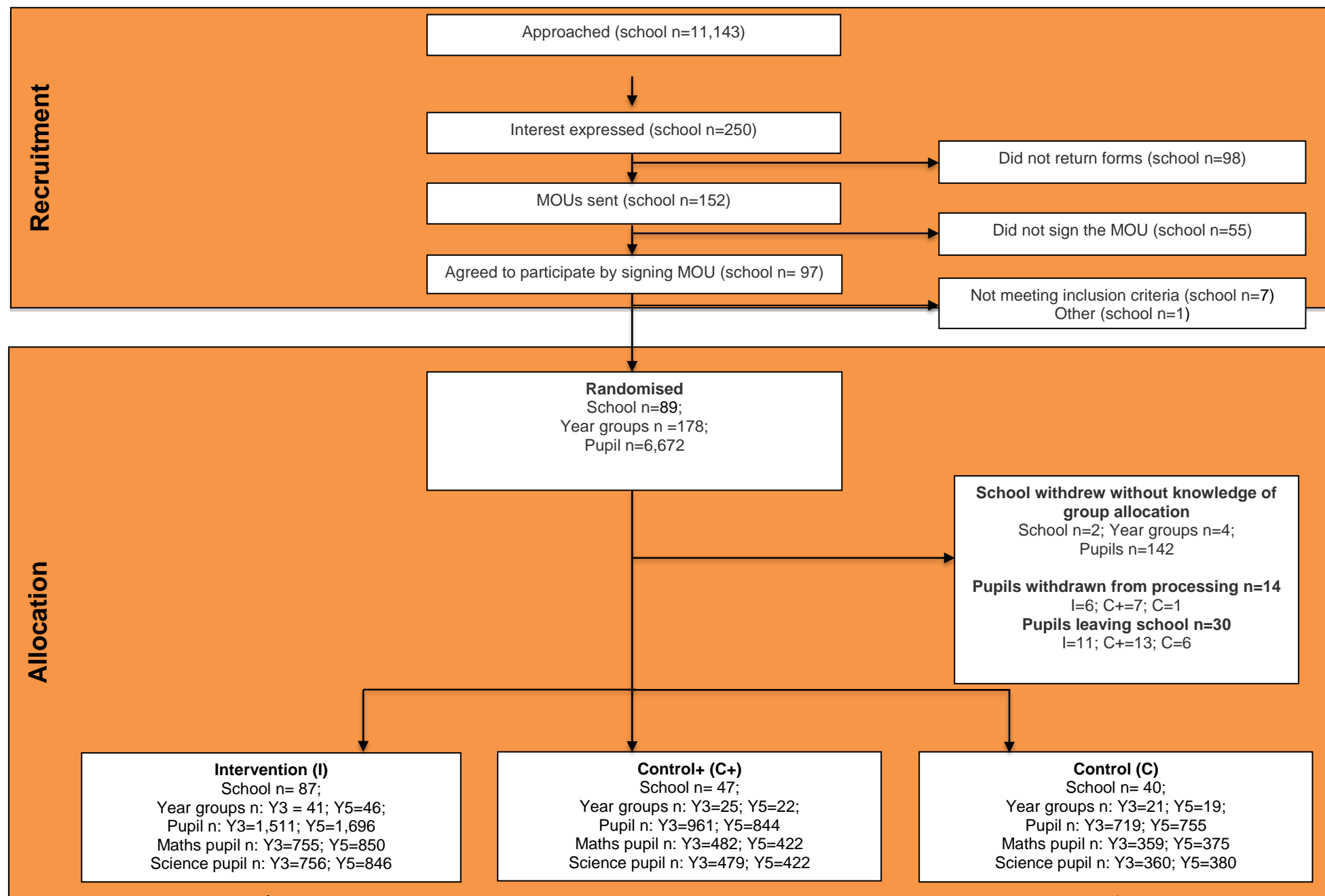
Participant

Figure 2 presents details of the participant flow through each stage of the trial. As described in the selection and recruitment section, Birkbeck College was responsible for school recruitment. It received MoUs from 97 schools. Of these, 89 were put forward for randomisation as five had only one class (either Year 3 or Year 5 or mixed class of Year 3 and Year 5), two did not provide pupil data, and one withdrew prior to randomisation. A further two schools withdrew from the trial after randomisation but prior to Birkbeck College informing them of their random allocation. We present these schools as unbiased dropout in the allocation window in Figure 2.

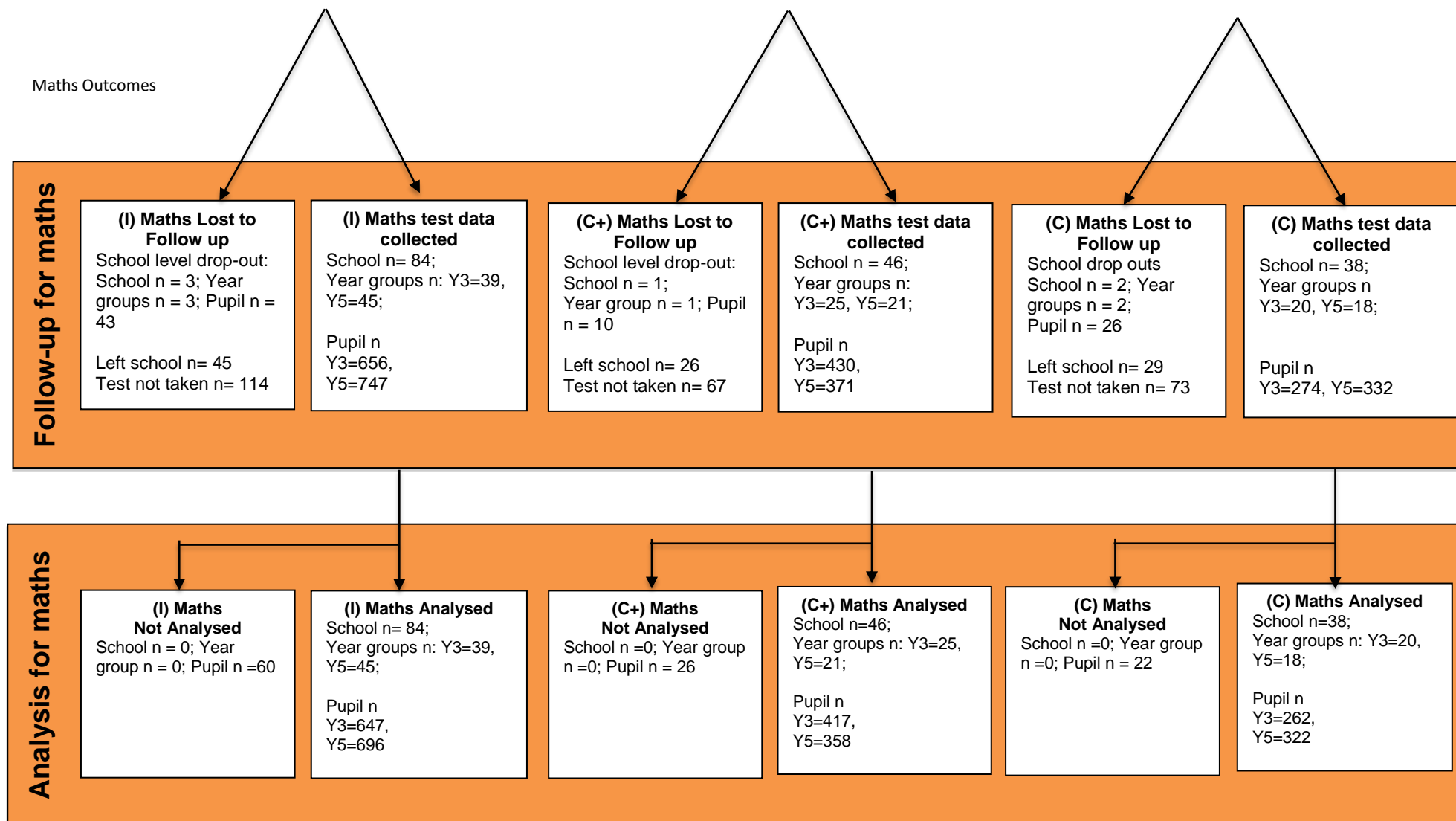
Schools provided administrative pupil data (pupil names, UPN, and date of birth) to Birkbeck College in the summer term of 2016/2017. In September and October 2017, schools sent an update on the pupil list to enable Birkbeck to remove pupils who had left school as well as those who had withdrew from data processing. These pupils were not included in the trial analysis. As this was prior to schools knowing their randomisation group, we present these numbers in the allocation window and consider them as unbiased dropouts. The numbers of schools, year groups, and pupils are presented in the allocation window broken down by randomisation groups. For the purposes of attrition, we will consider these numbers as the ones meant to be followed up in the trial.

As mentioned before, there were two primary outcomes for the trial, maths and science. Therefore, we present further stages of the participant flow diagram (follow-up and analysis) by the outcome measures. Page 2 of the flow diagram presents numbers of schools, year groups, and pupils followed up and analysed for the maths outcomes. Similarly, page 3 of the flow diagram presents numbers followed up and analysed for the science outcomes.

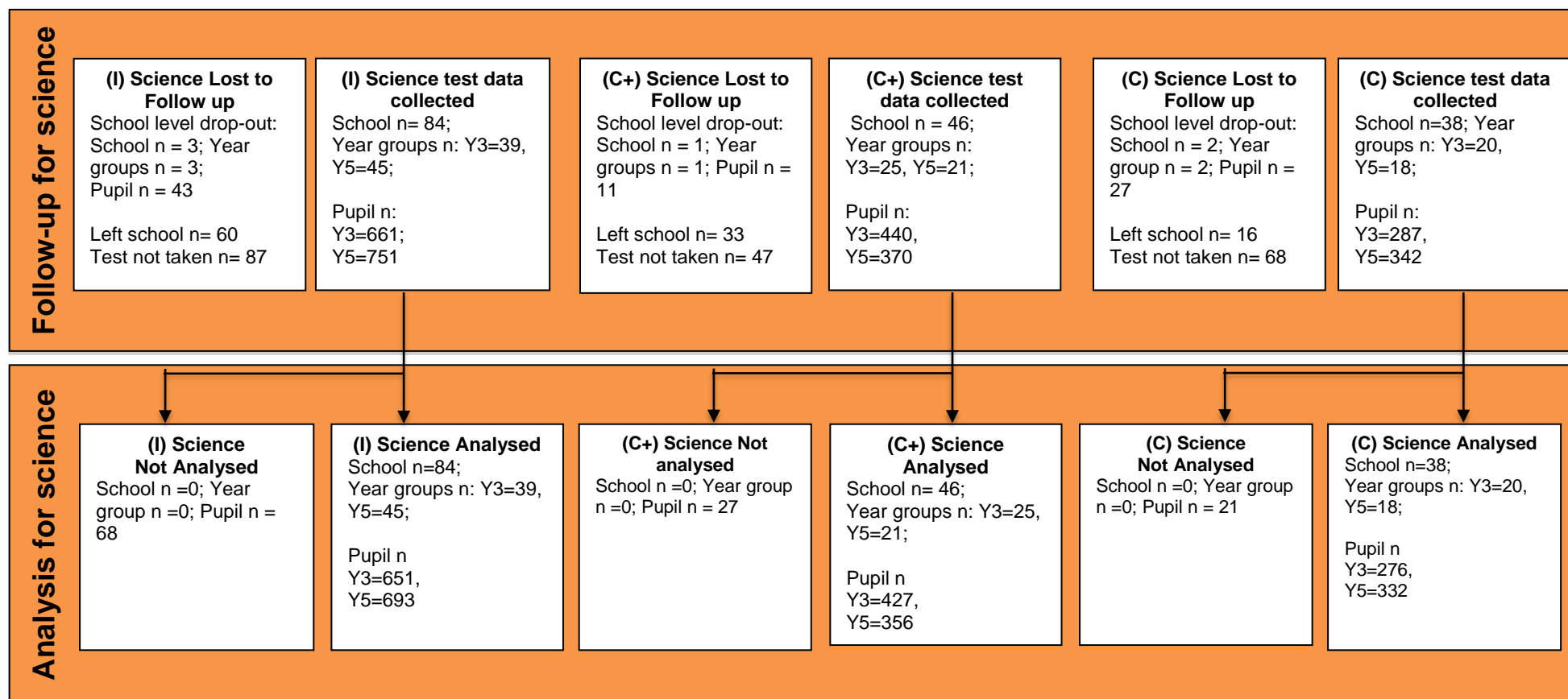
Figure 2: Participant flow diagram for Counterintuitive Concepts



Maths Outcomes



Science Outcomes



Attrition

In terms of attrition to measurement, we lost three schools that did not agree to primary outcomes tests. This meant that school-level attrition was relatively low at 3%. We also lost all the year groups from these schools and therefore year-group-level attrition was also at 3%. This resulted in a loss of 160 pupils from the follow-up testing. In addition to this, we also lost pupils from the analysis as they had either left the school prior to testing (n = 209), were absent on the day of the testing (n = 456), or we could not match their prior attainment data on the NPD (n = 224). Table 9 presents pupil-level attrition based on the numbers from the participant flow diagram. On average, we lost 17% pupils from maths analysis and 16% pupils from science analysis.

Table 9: Pupil level attrition from the trial—combined Y3 and Y5 figures

	Maths		Science		Total
	Intervention group	Combined control group	Intervention group	Combined control group	
Number of pupils meant to be followed up ¹²	1605	1638	1602	1641	6486
Number of pupils analysed	1343	1359	1344	1391	5437
Pupil level attrition	16%	17%	16%	15%	16.17%
Overall attrition	17%		16%		

Table 10 provides details of minimum detectable effect sizes at different stages in the trial. Up until the randomisation, we did not separate Year 3 and Year 5 models. Therefore, only one MDES per subject is presented. Since the outcome measures were different for each Year group, we ran maths and science models for each Year group separately. MDES at analysis is calculated based on the numbers in these models. In these models, maths and science outcomes of the intervention pupils were compared against the outcomes of the combined control groups (control group and control-plus group together).

¹² This includes number of schools and pupils retained in the trial after excluding schools that withdrew prior to the knowledge of group allocation and pupils that withdrew from data processing.

Table 10: Minimum detectable effect size at different stages

		Maths				Science			
		Protocol	Randomisation	Analysis (i.e. available pre- and post- test)		Protocol	Randomisation	Analysis (i.e. available pre- and post- test)	
				Year 3	Year 5			Year 3	Year 5
MDES		0.125	0.135	0.180	0.185	0.130	0.135	0.200	0.200
Correlation between pre-test (+other covariates) and post-test		0.65	0.65	0.58	0.51	0.65	0.65	0.54	0.53
Intraclass correlations (ICCs)		0.13	0.13	0.07	0.07	0.13	0.13	0.09	0.09
Blocking/stratification or pair matching		School blocking	School blocking	School blocking	School blocking	School blocking	School blocking	School blocking	School blocking
Alpha		0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
Power		0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
One-sided or two-sided?		2	2	2	2	2	2	2	2
Average cluster (class) size		20.25	18.47	10.96	10.58	20.25	18.47	11.19	10.62
Number of schools	intervention	100	86	39	45	100	86	39	45
	control	100	86	45	39	100	86	45	39
	total	100	86	84	84	100	86	84	84
Number of pupils	intervention	2025	1589	647	696	2025	1589	651	693
	combined control group	2025	1629	679	680	2025	1629	703	688
	total	4050	3218	1326	1376	4050	3218	1354	1381

Pupil and school characteristics

In total, 87 schools were involved in the trial. Three schools were lost to primary outcomes tests. Table 11 presents key baseline characteristics of the remaining 84 schools who were included in the primary analysis. As the randomisation was at Year group level, the schools had two of the three randomisation groups. Schools where Year 3 was randomised to intervention, Year 5 was randomised to one of the two control groups and these schools were compared against the schools where Year 3 was randomised to one of the two control groups and Year 5 was randomised to intervention. Characteristics of the two types of schools are presented in Table 11. Looking at the table, there is a small imbalance between the characteristics of the schools in school governance, school Ofsted rating, and whether schools are urban or rural.

Table 11: Baseline comparison for analysed groups (school characteristics)

School-level (categorical)	Schools with Year 3 intervention group and Year 5 control group		Schools with Year 3 control group and Year 5 intervention group	
	n/N (missing)	Count (%)	n/N (missing)	Count (%)
School Governance				
Academy or Free School	14/39 (0)	36%	14/45 (0)	31%
Maintained	25/39 (0)	64%	31/45 (0)	69%
Ofsted rating				
Outstanding	4/39 (0)	10%	7/45 (0)	15%
Good	30/39 (0)	77%	35/45 (0)	78%
Requires improvement	3/39 (0)	8%	3/45 (0)	7%
Inadequate	2/39 (0)	5%	0/45 (0)	0%
Urban or rural				
Urban	26/39 (0)	67%	32/45 (0)	71%
Rural	13/39 (0)	33%	13/45 (0)	29%
Primary school type				
Primary/Combined	37/39 (0)	95%	43/45 (0)	96%
Junior	2/39 (0)	5%	1/45 (0)	2%
Other type	0/39 (0)	0%	1/45 (0)	2%
Percentage pupils FSM-eligible 2016/2017 (5 point scale)				
Lowest 20%	6/39 (0)	15%	7/45 (0)	16%
2 nd lowest 20%	9/39 (0)	23%	11/45 (0)	24%
Middle 20%	8/39 (0)	21%	5/45 (0)	11%
2 nd highest 20%	10/39 (0)	26%	11/45 (0)	24%
Highest 20%	6/39 (0)	15%	11/45 (0)	24%
Form Entry				
One form – Year 3 and 5	21/39 (0)	54%	26/45 (0)	58%
Two form – Year 3 and 5	10/39 (0)	26%	11/45 (0)	24%
Three form – Year 3 and 5	2/39 (0)	5%	3/45 (0)	7%
Mixed – all other form entry	6/39 (0)	15%	5/45 (0)	11%
School-level (continuous)	n (missing)	Mean% (SD)	n (missing)	Mean% (SD)
% FSM 2016/2017	39 (0)	15% (13)	45 (0)	16% (12)

Table 12 presents characteristics of pupils who were included in the primary analysis. The table presents FSM eligibility and pre-test scores (EYFSP) for pupils included in each of the four models. In addition to this, we also calculated the baseline effect size using the EYFSP data for analysed groups. As seen in the table, the effect size confidence intervals straddle zero which suggests no evidence of a difference in EYFSP scores of the two randomisation groups. Appendix H presents the distribution of pre-test results by analysed groups.

Table 12: Baseline comparison for analysed groups (pupil characteristics)

	Intervention group		Combined control groups		
Pupil level (categorical)	n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Eligible for FSM (Ever, Spring 2018)					
Maths Year 3	179/647 (0)	28%	202/679 (0)	30%	
Maths Year 5	246/696 (0)	35%	198/680 (0)	29%	
Science Year 3	175/651 (0)	27%	202/703 (0)	29%	
Science Year 5	245/693 (0)	35%	197/688 (0)	29%	
Pupil level (continuous)	n (missing)	Mean (SD)	n (missing)	Mean (SD)	Effect Size Hedges g (95%CI)
Foundation Stage Profile Score					
Maths Year 3	647 (0)	2.02	679 (0)	2.00	-0.0015 (-0.20, 0.20)
Maths Year 5	696 (0)	1.93	680 (0)	1.94	-0.09 (-0.32, 0.13)
Science Year 3	651 (0)	2.00	703 (0)	1.99	0.01 (-0.17, 0.19)
Science Year 5	693 (0)	1.84	688 (0)	1.92	0.01 (-0.23, 0.26)

Outcomes and analysis

Primary ITT analysis

We present the score distribution for four outcome measures by the intervention group and the combined control groups in Appendix I. The maximum possible score range for each of the tests was 55 for PTM8, 65 for PTM10, 40 for PTS8, and 50 for PTS10. As seen in the histograms, most distributions are approximately normal with an exception that in the maths outcomes, the intervention groups tend to more than one peak (mode).

Table 13 presents findings from the main analyses. As described in the methods section, we ran two separate models for each subject—Year 3 and Year 5. Effect sizes from these models constitute secondary analysis. The primary analysis for each subject was a combined effect size across two year groups (presented in the last column of Table 13). As seen in the table, the combined effect size for the primary analysis in maths was 0.09 (-0.01, 0.19). As the confidence intervals straddle zero, we cannot reject the null hypothesis. This means that the statistical evidence does not meet the threshold to conclude that the true impact of Stop and Think was non-zero. The combined effect size for the primary analysis in science was 0.12 (0.02, 0.22), which does not straddle zero. Therefore, we can reject the null hypothesis. This result suggests that the intervention had a positive effect on pupils' science attainment when the combined Year 3 and Year 5 science results are taken into consideration. On average, intervention pupils scored higher in science when compared to the control group pupils. Table 15 presents the parameters used in estimating the effect size for each model.

Table 13: Outcomes analyses, GL Assessment PTM and PTS—primary, secondary, and FSM-only analyses

	Raw means				Effect size			Primary analysis: combined effect size Y3 and Y5 (95% CI)
	Intervention group		Control group					
Outcome	N (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges g (95% CI) (secondary analysis)	p-value	
PTM8 GL test score Maths Year 3	656 (70)	25.7 (24.8, 26.6)	704 (122)	25.1 (24.2, 25.9)	1326 (647; 679)	0.03 (-0.12, 0.18)	0.67	0.09 (-0.01, 0.19)
PTM10 GL test score Maths Year 5	747 (89)	31.3 (30.3, 32.3)	703 (73)	29.7 (28.7, 30.8)	1376 (696; 680)	0.14 (-0.002, 0.28)	0.05	
PTS8 GL test score Science Year 3	661 (66)	23.2 (22.7, 23.7)	727 (97)	22.7 (22.3, 23.2)	1354 (651; 703)	0.07 (-0.08, 0.22)	0.34	0.12 (0.02, 0.22)
PTS10 GL test score Science Year 5	751 (81)	29.3 (28.7, 29.8)	712 (67)	28.4 (27.8, 29.0)	1381 (693; 688)	0.17 (0.03, 0.32)	0.02	
PTM8 GL test score (FSM only) Maths Year 3	181 (28)	21.2 (19.5, 22.8)	210 (38)	20.9 (19.4, 22.4)	381 (179; 202)	0.19 (-0.02, 0.40)	0.07	
PTM10 GL test score (FSM only) Maths Year 5	260 (42)	26.5 (24.9, 28.1)	208 (33)	24.1 (22.4, 25.8)	444 (246; 198)	0.16 (-0.04, 0.36)	0.11	
PTS8 GL test score (FSM only) Science Yeas 3	176 (21)	20.2 (19.3, 21.2)	208 (27)	20.8 (20.0, 21.6)	377 (175; 202)	0.01 (-0.19, 0.20)	0.96	
PTS10 GL test score (FSM only) Science Year 5	262 (34)	26.0 (25.2, 26.9)	203 (26)	25.4 (24.5, 26.4)	442 (245; 197)	0.10 (-0.13, 0.33)	0.39	

Table 14: Secondary outcomes analyses: Chimeric Animal Stroop task

	Raw means				Effect size			Primary analysis: combined effect size Y3 and Y5 (95% CI)
	Intervention group		Control group					
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (intervention; control)	Hedges g (95% CI) (secondary analysis)	p-value	
Chimeric Animal Stroop task Year 3	1256 (0)	10.3 (10.1, 10.5)	1403 (0)	10.3 (10.1, 10.4)	2659 (1256; 1403)	-0.01 (-0.13, 0.10)	0.84	0.03 (-0.05, 0.11)
Chimeric Animal Stroop task Year 5	1437 (0)	13.2 (13.0, 13.3)	1354 (0)	12.9(12.7, 13.1)	2791 (1437; 1354)	0.08 (-0.05, 0.20)	0.23	

Table 15: Effect size estimation

Outcome	Unadjusted differences in means	Adjusted differences in means	Total variance from a model without covariates	Population variance (if available)
PTM8 GL test score Maths Year 3	0.6	0.36	131.96	
PTM10 GL test score Maths Year 5	1.6	1.95	200.56	
PTS8 GL test score Science Year 3	0.5	0.44	39.27	
PTS10 GL test score Science Year 5	0.9	1.36	60.39	

Missing data analysis

As described in the methods section, we explored the association of missingness with observable school and pupil variables with regard to each of the models noted below. In each case, the probability that the outcome measure was missing (compared to observed) was modelled using a multilevel logistic model.

- In the Year 3 maths model, we should have 1,552 pupils with PTM8 scores; 51 pupils had missing EYFSP score and 226 had missing PTM8 scores. The probability that the outcome measure was missing was found to be significantly associated with (a) lower than average prior attainment, (b) a pupil from a school with a three-form entry, and (c) pupils from junior schools compared to (combined) primary schools.
- In the Year 5 maths model, we should have 1,612 pupils with PTM10 scores; 92 pupils had missing EYFSP score and 236 had missing PTM10 scores. The probability that the outcome measure was missing was found to be significantly associated with (a) lower than average prior attainment and (b) everFSM status of the pupils.
- In the Year 3 science model, we should have 1,551 pupils with PTS8 scores; 44 pupils had missing EYFSP score and 197 had missing PTS8 scores. The probability that the outcome measure was missing was found to be significantly associated with (a) lower than average prior attainment, (b) a pupil from a school with a three-form entry (c) pupils from rural schools, (d) pupils from junior schools compared to (combined) Primary schools, and (e) pupils from schools with missing data on Key Stage 2 science attainment.
- In the Year 5 science model, we should have 1,611 pupils with PTS10 scores; 99 pupils had missing EYFSP score and 230 had missing PTS8 scores. The probability that the outcome measure was missing was found to be significantly associated with pupils from schools that belonged to the lowest performance band in Key Stage 2 science.

These patterns of missing data demonstrate that the data was not missing completely at random (MCAR).

Missing data was imputed (with chained equations, implemented using the MICE package in R) under the assumption that data was missing at random (MAR). We ran four different imputation models. Each model included the primary outcome variable, prior attainment EYFSP score, randomisation stratification (form-entry set-up), intervention or combined control group, urban or rural school, school Ofsted rating, school type, and everFSM eligibility.

The main ITT models were run using each of the imputed datasets for each subject separately for Year 3 and Year 5. The results from the imputed datasets were pooled to give coefficients and standard errors that took account of the imputation variance. For Year 3 maths, the complete data analysis gave the coefficient of being in the intervention as 0.39 (-0.44, 1.21). This compares to a completers model raw intervention coefficient of 0.36 (-1.31, 2.03). For Year 5 maths, the complete data analysis gave the coefficient of being in the intervention group as -0.06 (-0.13, 0.02). This compares to the completers model with raw intervention coefficient of 1.947 (-0.002, 3.897). For Year 3 science, the complete data gave the coefficient of being in the intervention group as -0.02 (-0.08, 0.04). This compares to the completers model with raw intervention coefficient of 0.44 (-0.46, 1.35). For Year 5 science, the complete data gave the coefficient of being in the intervention group as 1.50 (0.52, 2.48), which is compared to the completers model with raw coefficient of 1.36 (0.24, 2.47). These results from the imputed models imply that even with the imputed values for the

missing data, the results were fairly consistent with the ITT models and we could be certain that the completers analyses are unlikely to be biased.

Complier Average Causal Effect (CACE)

As described in the methods section, Birkbeck College provided us with the number of Stop and Think sessions that each class had experienced. From this, it was assumed that all pupils within a given class would have received the same number of Stop and Think sessions. The compliance data suggests that a majority of intervention pupils had received up to 30 sessions with a small proportion experiencing up to 32 sessions; 8% of intervention pupils did not engage with the Stop and Think sessions at all and other 8% received 30 or more sessions (maximum compliance); 14% experienced one to ten sessions (low compliance), 15% experienced 11 to 20 sessions (medium compliance), and 63% experienced more than 21 sessions (high compliance). Further details are included in Appendix K. As the delivery team derived this data directly from the schools' computer systems, we used all the data as provided.

Results from the models with the compliance measure suggested that there was no evidence that the number of Stop and Think sessions was associated with Year 3 pupils' attainment in maths or science. In addition, the number of Stop and Think sessions was not associated with Year 5 pupils' maths attainment. However, the effect size for the Year 5 science outcome was 0.0069 (0.0002, 0.0136) which suggests that the Year 5 pupils with higher number of Stop and Think sessions performed better in PTS10 compared to the control group pupils.

Secondary outcome analyses

The combined effect sizes for the primary analyses were based on four individual outcome measures. The individual year group outcome models constituted the secondary analyses. Effect sizes and confidence intervals for these models are also presented in Table 13. As seen in the table, the confidence intervals straddle zero for three of the four secondary analyses models. This means the statistical evidence does not meet the threshold to conclude that the true impact of Stop and Think on Year 3 pupils' maths attainment, Year 5 pupils' maths attainment and Year 3 pupils' science attainment was non-zero. The Year 5 science model results were statistically significant at $p < 0.05$. This means that the intervention pupils achieved, on an average, higher scores in PTS10 compared to the control group pupils with an effect size of 0.17 (0.03,0.32).

The outcomes of the Chimeric Animal Stroop task were analysed via multilevel linear regression models. We checked the data for normality and the histograms with distribution are presented in Appendix J. As detailed in the methods section, we ran two models- one for each Year group; these are summarised in Table 14. Looking at the effect sizes- the combined effect size across the two Year groups and the effect size from each individual Year group model suggested that the confidence intervals straddled zero. This means the statistical evidence does not meet the threshold to conclude that the true impact of Stop and Think on pupils' inhibitory control was non-zero.

Additional analyses

As mentioned in the methods section, we conducted two additional analyses. The first analysis looked at attainment differences between the intervention group and the control-plus group (Table 16). The combined effect sizes for maths and science were 0.13 (0.002, 0.25) and 0.15 (0.02, 0.27) respectively. These results suggest that the pupils who received Stop and Think, on average, scored higher on PTM and PTS when they were compared with the pupils who received See+. The results demonstrate that the Stop and Think programme had an impact on pupils' maths and science attainment over and above a similar computer programme. Looking at each year group model separately (Table 16), the evidence suggest that the intervention had no statistically significant effect on Year 3 pupils' maths or science attainment (over and above See+). But, the effect sizes for the Year 5 maths and science models were 0.22 (0.05, 0.39) and 0.23 (0.05, 0.41) respectively which suggests that Year 5 pupils who received Stop and Think scored higher on PTM10 and PTS10 compared to those who received See+. This difference was statistically significant at $p < 0.05$.

Table 16: Additional analysis—intervention versus control-plus

Outcome	n in model (intervention; control-plus)	Hedges g (95% CI)	p-value	Combined effect size (Y3 and Y5, 95% CI)
PTM8 GL test score Maths Years 3	1064 (647; 417)	0.02 (-0.17, 0.20)	0.85	0.13 (0.002, 0.25)
PTM10 GL test score Maths Years 5	1054 (696; 358)	0.22 (0.05, 0.39)	0.01	
PTS8 GL test score Science Years 3	1078 (651; 427)	0.07 (-0.11, 0.25)	0.43	0.15 (0.02, 0.27)
PTS10 GL test score Science Years 5	1049 (692; 356)	0.23 (0.05, 0.41)	0.01	

In the second additional analysis, which is presented in Table 17, we looked at attainment differences between the control-plus group pupils (See+) and the business-as-usual control group pupils to explore the impact of the See+ computer programme on pupils' maths and science attainment. The combined effect size for maths was -0.08 (-0.23, 0.06). Although the effect size is negative and seemingly favouring the control group pupils, this difference was not statistically significant. This means the difference could have arisen by chance. Individual models for Year 3 and Year 5 maths also suggested that there was no evidence that the See+ programme had an impact on pupils' maths attainment when compared with the control group pupils. The picture was very similar when science attainment was considered. The combined effect size for science was -0.08 (-0.22, 0.06) which suggest that there is no evidence that the See+ programme had an effect on pupils' science attainment compared to pupils in the business-as-usual control group. Similarly, the separate models for each Year group also demonstrated that there is no evidence that the See+ programme had any effect on Year 3 or Year 5 pupils' science attainment when compared with the control group pupils.

Table 17: Additional analysis—control-plus versus control

Outcome	n in model (control-plus; control)	Hedges g (95% CI)	p-value	Combined effect size (Y3 and Y5, 95% CI)
PTM8 GL test score Maths Years 3	679 (417; 262)	0.02 (-0.19, 0.23)	0.84	-0.08 (-0.23, 0.06)
PTM10 GL test score Maths Years 5	680 (358; 322)	-0.19 (-0.40, 0.02)	0.08	
PTS8 GL test score Science Years 3	703 (427; 276)	-0.04 (-0.24, 0.17)	0.72	-0.08 (-0.22, 0.06)
PTS10 GL test score Science Years 5	688 (356; 332)	-0.12 (-0.33, 0.09)	0.24	

Subgroup analyses

We conducted separate analyses for a subset of pupils who received FSM at some point in the previous six Years (everFSM) only. Effect sizes for each of the four models are presented in Table 12. As seen in the table, the confidence intervals straddle zero for each effect size, which suggests that the statistical evidence does not meet the threshold to conclude that the true impact on everFSM pupils' maths or science attainment was non-zero.

Results from the interaction models are summarised in Table 17. In these models, everFSM and gender were interacted with the intervention term respectively. We ran four separate models for each variable of interest to explore the differential impact of the intervention. These results suggest that the intervention did not have a statistically significant differential effect on maths or science attainment when pupil everFSM status and gender were considered. This means the intervention did not have a differential impact for boys compared to girls or pupils with everFSM status compared to those who were not everFSM.

Table 18: Results of interaction models

Outcome variable	Variable of interest	Raw interaction coefficient	Standard error	p-value
Year 3 maths	Eligible for FSM (EverFSM6, Spring 2018)	1.43	1.17	0.22
Year 5 maths	Eligible for FSM (EverFSM6, Spring 2018)	0.32	1.43	0.83
Year 3 science	Eligible for FSM (EverFSM6, Spring 2018)	-0.46	0.65	0.47
Year 5 science	Eligible for FSM (EverFSM6, Spring 2018)	-0.91	0.78	0.24
Year 3 maths	Gender	-1.67	0.98	0.09
Year 5 maths	Gender	-0.64	1.24	0.61
Year 3 science	Gender	0.53	0.56	0.35
Year 5 science	Gender	-0.40	0.68	0.56

Cost

The average cost of Stop and Think was £5.76 per pupil per year when averaged over three years. This estimate is based on the delivery of the intervention to one year group. It is estimated on the basis of the programme software being free, and includes costs of the initial training and ongoing support from

Birkbeck provided in this trial for the first year only. The assumption is that schools could just use the handbook for the subsequent two years without training. This estimate does not include costs associated with staff time such as training and preparation, Birkbeck College developed a beta version of the Stop and Think computer programme during the development phase. This version was used for the evaluation. The cost of participating in this evaluation was covered by a grant from the Education Endowment Foundation and the Wellcome Trust.

In the online survey, we asked teachers to provide us with the costs relating to obtaining additional resources, if any. Of the 61 Stop and Think teachers that responded to the online surveys, 53 said that their school did not require any additional resources to run Stop and Think, seven respondents were not sure, and one respondent said that they required new hardware.

The activities and time involved in delivering Stop and Think, as reported by teachers, were as follows:

- Training: While not all teachers received the training (nine did not), over half of the respondents (32) said that the training lasted between 16 and 30 minutes.
- Preparing for the first Stop and Think session: All but four respondents said that preparing for their first session took less than 15 minutes with over half of respondents (32) reporting it took under five minutes. In addition, five teachers did not spend any time preparing for their first session.
- Preparing for each Stop and Think session: The majority of respondents (42) indicated that it took between one and five minutes to prepare for each session, excluding time to log into the system.
- Setting up, including logging in to the software: Almost all of the teachers (58) reported that it took between one and five minutes to set up and log in to the software.
- Delivering each session:
 - over a third (27) of respondents indicated that it took between six and 15 minutes to deliver Stop and Think; roughly another third (23) estimated it took between 16 to 20 minutes; and
 - nine respondents indicated that it took between 21 and 30 minutes—supporting the interview evidence that there were issues with the timeout function of the software, which should automatically time-out after 12 minutes.

Overall, the time involved in preparing for, and setting up, Stop and Think was less than five minutes. However, it should be noted that five minutes is a large proportion of time spent to setup a session which is meant to be twelve minutes long. On average, the training lasted for 23 minutes and it took less than 15 minutes to prepare for the first Stop and Think session. On average, teachers delivered the Stop and Think programme for 16 minutes.

Implementation and process evaluation

Implementation

Research question: Have schools implemented the intervention in the way it was intended?

- The Stop and Think interactive computer programme asks questions and indicates whether the answers entered are right or wrong. The intervention's requirement was that Stop and Think should be run as a 15-minute session with Year 3 or Year 5 pupils before maths or science lessons three times a week for ten weeks (30 sessions in total). A classroom teacher facilitates the sessions.
- Implementation of Stop and Think was often challenging for over half of the teachers surveyed who had experienced issues in using the software programme, which caused delays and impeded the smooth running of the sessions.
- Where there were mixed-age classes in two of the case-study schools, some pupils were waiting for the maths or science lesson to begin while Year 3 or Year 5 pupils participated in the Stop and Think sessions.

Fidelity

Research question: Have schools implemented the intervention in the way it was intended?

- Stop and Think was delivered as intended by a majority of teachers (44 out of 61) surveyed.
- Fitting Stop and Think into the school timetable was a challenge because the curriculum time was already squeezed by other competing demands. The majority of teachers did not endorse the roll-out of the programme to other schools (47 out of 61) because of difficulty in fitting delivery into the school day, software problems, pupil engagement, accuracy of content, quality of animation, and the content being too easy.
- Teachers were positive about the training and Teacher Guide they had received, which they thought had prepared them to deliver Stop and Think. Most indicated that their school did not require any additional resources to run the programme. They indicated that Birkbeck gave them a good level of support.

Suitability

Research question: Is the intervention appropriate for pupils of this age and in these lessons?

- A majority of the teachers surveyed considered that the Stop and Think content was appropriately aligned with the curriculum for science and was suitable for their class.
- While half of the teachers thought that Stop and Think was suitable for their class for maths, just under half thought it was too easy.
- Most teachers indicated that Stop and Think was suitable for pupils in upper Key Stage 1 (Year 2) and Key Stage 2. The programme was also suitable for pupils with SEN.

Outcomes

Research question: is the intervention appropriate for pupils of this age and in these lessons?

- A majority of teachers thought that Stop and Think had had a positive impact on the mathematical ability and science ability of pupils in their classes. Other impacts of using the programme were said to be on pupils taking time to consider their response before answering questions, developing numeracy and science skills, enhanced confidence, improving engagement in learning, and developing social skills such as listening and considering other pupils' points of view.

- A majority of teachers considered that using Stop and Think had had a positive impact on them as a class teacher. The impacts included developing a better understanding of how pupils learn and gaining more insights into pupils' reasoning.

Formative findings

Research question: Can programme materials and delivery be improved for the future?

- The survey evidence suggests that Stop and Think could be improved if it were linked better to the teaching cycle and pupils' learning activities. This would mean the programme being more flexible and offering teachers more control so that they could use it to refer to topics already covered by their class. Providing this function would eliminate the situation where the programme asks questions on topics the pupils have not covered yet, which can cause confusion and present a challenge for teachers to hastily pre-teach a topic. Teachers thought that Stop and Think sessions were more successful where the content had been covered in the curriculum.
- The Stop and Think programme would benefit from having more advanced animations and graphics (projected onto whiteboards) which would gain and maintain the interest of pupils and engage them more.
- Addressing the software issues in delivering Stop and Think would help to make using the programme a more productive experience for teachers and pupils.
- The inclusion of harder questions, especially in maths, would make Stop and Think a more appropriate resource for use in schools.

Stop and Think— findings from the survey and interviews

The findings from the survey and interviews with teachers delivering Stop and Think are presented below.

Fidelity

Research question: Have schools implemented the intervention in the way it was intended?

The teacher survey explored fidelity in the delivery of the Stop and Think intervention. The survey found that 43 of the 61 teacher respondents indicated that their class received Stop and Think for ten weeks and 44 indicated that their class received Stop and Think three times a week. While 35 teachers indicated that their class received Stop and Think at the start of a science or maths lesson, 18 more indicated that this happened 'sometimes'.

It is worth noting that there was some variation in the type of delivery of the intervention. For example, in the four Stop and Think sessions we observed as part of the case studies, the sessions were a whole-class activity. There was no paired or small group work. After each question in the programme, the teacher selected one pupil to give an answer and then asked the rest of the class whether they agreed or not with the pupil's answer by raising their hands. In three of the schools, the teacher inputted the answer given by the majority of pupils on the computer and in the other school pupils came up and inputted the answer either by touch screen or using the mouse. In each case, the teacher facilitating the session did not influence pupils by saying whether he/she thought the answer voted for by the pupils was right or wrong, though sometimes the teacher asked pupils to explain the thinking underlying the answer they had given. The teacher's involvement was focused on keeping the session moving. Most pupils were engaged in the sessions we observed.

When we interviewed teachers in ten of the 87 schools participating in the trial, some said that it was a challenge to fit Stop and Think into the school timetable. They elaborated that the curriculum was already squeezed and there were competing demands on time, one remarking that 'ten minutes is

precious for learning timetables and spelling'. Other comments revealed that schools could not always do Stop and Think three times a week owing to having a full teaching agenda. Therefore, it was difficult to commit to the right amount of time every week and schools could not always do Stop and Think before maths and science lessons, so sometimes they ran the programme first thing in the morning or after the lunch break.

In addition to the data collection for the process evaluation, Birkbeck College collected data from schools' computer systems directly. This data included the number of Stop and Think sessions each class (and therefore a pupil within a given class) experienced. Although this data provided a finer measure of compliance at pupil level, it did not account for pupil absence—missing a session. This compliance data suggested that the level of compliance was moderate to high where 63% of intervention pupils experienced more than 21 sessions (high compliance), 15% experienced 11 to 20 sessions (medium compliance), 14% experienced 1 to 10 sessions (low compliance), and 8% intervention pupils did not engage with Stop and Think sessions at all (no compliance). It is also worth noting that 8% of pupils experienced full compliance (that is, at least 30 sessions).

Suitability

Research question: Is the intervention appropriate for pupils of this age and in these lessons?

Suitability of curriculum

The survey asked teachers about the suitability of the programmes in terms of content, subject matter, and age-appropriateness. The survey found that around half (30) of the teachers considered that the content (subject matter) of Stop and Think was pitched appropriately for their class for maths, 25 considered the content was 'too easy', and two considered it 'too difficult.' A majority of teachers (49) considered that the Stop and Think content (subject matter) was appropriately aligned with the curriculum for maths. The survey found that 45 teachers considered that the content (subject matter) was pitched appropriately for their class in science with four saying it was 'too easy' and five saying it was 'too difficult'. A majority of teachers (50) considered that the Stop and Think content (subject matter) was appropriately aligned with the curriculum for science.

When we interviewed teachers, some noted that there was a mismatch between the order of Stop and Think topics covered in the programme and the order of topics covered by the curriculum. This was exemplified by a Stop and Think activity focusing on 'light' in a school where pupils had not, as yet, covered 'light' in their science lessons; on another occasion, a teacher had to pre-teach fractions in a maths class to enable pupils to understand and answer the Stop and Think questions on fractions. Where the programme referred to topics covered in the curriculum, teachers thought it was useful for embedding learning because it gets them to think about the topic and related concepts. They said that Stop and Think sessions were more successful where the content had been covered in the curriculum. Another observation was that the maths and science topics in Stop and Think should be delivered separately, explaining, for example, that it would be more helpful to progress from a maths activity to a maths lesson.

Suitability for pupils in year groups

The survey investigated whether teachers thought the intervention was suitable for pupils in the year groups. This was important as the intervention was targeted at pupils in Year 3 and in Year 5 in Key Stage 2. We were unsure of the suitability of the intervention for these age groups so included relevant questions in the survey. The survey asked teachers whether Stop and Think was suitable for pupils in upper Key Stage 1 (Year 2). Most teachers considered that the programme was suitable: 13 'to a great extent', 21 'to some extent', and nine 'to a little extent'. Only two teachers considered that the programmes were not suitable. The survey asked teachers whether Stop and Think was suitable for pupils in lower Key Stage 2 (Years 3 and 4). Most teachers considered that the programmes were

suitable: 22 'to a great extent', 22 'to some extent', and six 'to a little extent'. Only two teachers considered that the programmes were not suitable. The survey asked teachers whether Stop and Think was suitable for pupils in upper Key Stage 2 (Years 5 and 6). Most teachers considered that the programmes were suitable: 13 'to a great extent', 22 'to some extent', and 16 'to a little extent'. Only five teachers considered that the programme was not suitable.

Some teachers we interviewed pointed out that using Stop and Think with mixed-age classes in small schools was a challenge. For example, where classes included Years 4, 5, and 6 pupils, some of the pupils not involved were waiting for the maths or science lesson to begin. Commenting on the maths and science content of the programme, some teachers thought that the maths content was pitched too low and the questions were too easy with the result that some pupils became frustrated because they were not learning anything. In contrast, teachers said that the science questions were quite hard and particularly difficult for pupils where the topics had not been covered to date in the curriculum.

Suitability for pupils with SEN

The survey asked teachers whether Stop and Think was suitable for pupils with special educational needs (SEN). Most survey respondents considered that the programme was suitable: six 'to a great extent', 34 'to some extent', and eight 'to a little extent'. Only four teachers considered the programme was not suitable.

When we interviewed teachers, some said that they read out the Stop and Think questions to the class because some pupils struggled with reading. They thought that the programme was suitable for pupils with SEN, noting, however, that some of them might not get as much out of the sessions as other pupils.

Suitability of programme questions and graphics

The views of teachers captured through the survey and case studies were that some pupils' engagement with Stop and Think was low because it was repetitive which made them lose interest and become less engaged with the session. The programme's questions and graphics were projected on to whiteboards for whole-class sessions and did not always gain and maintain pupils' interest and keep them engaged. Some teachers said that the programme's graphics were not advanced enough and that the programme should focus more on questions and have fewer graphics. Teachers noted that some older pupils in their class found the programme tedious because the maths questions were too simple.

Software issues

The survey investigated whether teachers had experienced issues with the software (for example, slow to load, screen freezing) during the delivery of Stop and Think sessions. Over half (38 of the 61 respondents) indicated that they had experienced problems with the software : 20 said 'sometimes', 13 'often', and five 'always'. In contrast, 11 teachers said 'seldom' and 12 said 'never'. Where teachers had experienced problems, they were asked to what extent they thought this impacted on the pupils' ability to engage fully with the Stop and Think programme. Most thought there had been an impact on pupils: nine 'to a great extent', 17 'to some extent', and 17 'to a little extent'.

Interviewed teachers said that they had had initial technical issues in loading the programme onto the laptop which they used to run the programme, sharing it with the class via a whiteboard, and to enter the answers to the questions in the programme. This was very frustrating and in some cases had affected their colleagues' engagement with using the programme. Where schools had used an interactive whiteboard this had worked successfully though sometimes loading the programme was slow. In the three of the four observed Stop and Think sessions we conducted, there were practical issues: loading the programme was slow and the programme was slow to move from one question to the next, with up to a one-minute pause between each question.

Training and resources

The survey asked for teachers' views on the training and resources associated with the programme. When asked whether the training provided by Birkbeck staff was suitable for preparing them to deliver Stop and Think, most (50) of the teachers considered that the training was suitable: 22 said 'highly suitable' and 28 said 'suitable'. Only four teachers said the training was 'not very suitable' (seven teachers had not received the training). The survey did not ask why teachers had not received training. Although it is difficult to assess accurately how not receiving training affected implementation fidelity, it is worth noting that the training focused mainly on the technical set-up and running of the programme and that teachers also had access to the written Teacher Guide. The survey findings were corroborated by the feedback teachers gave in interviews. They thought that the training provided was effective and that Birkbeck gave them a good level of support.

When asked how good the written Teacher Guide was in supporting them to deliver Stop and Think, over half gave a positive response: 14 said 'very good' and 26 said 'good', while 13 considered it 'acceptable'. Most of the teachers surveyed (53) indicated that their school did not require any additional resources to run Stop and Think. The teachers we interviewed thought that the written guidance was very clear and helped them to get started in using the programme.

Perceptions of impact

The survey asked teachers to what extent they thought the Stop and Think programme had a positive impact on the maths ability of pupils in their class. A majority indicated that there had been a positive impact: one said 'to a great extent', 22 said 'to some extent', and 30 said 'to a little extent'. Seven teachers indicated that there had been no impact on their pupils' maths ability.

The survey also asked teachers about the extent to which Stop and Think had a positive impact on the science ability of the pupils in their class. A majority indicated that there had been a positive impact: nine said 'to a great extent', 26 said 'to some extent', and 19 said 'to a little extent'. Six teachers indicated that there had been no impact on their pupils' science ability. When asked about a list of other impacts of Stop and Think on their pupils, 39 teachers reported an impact on pupils taking time to consider their response before answering questions, 30 reported an impact on developing science skills, 20 reported enhanced confidence, 18 reported improving engagement in learning, and 18 reported developing numeracy skills.

The survey asked teachers about the extent to which Stop and Think had a positive impact on them as a class teacher. A majority (44) indicated a positive impact: three said 'to a great extent', 21 said 'to some extent' and 20 said 'to a little extent'. Seventeen teachers indicated that the programmes had not had a positive impact on them. Examples of positive impact expressed by teachers in open-ended questionnaire responses were:

'It allowed me to develop my understanding of how the children in my class learn and to analyse what they know, how clearly they understand concepts and to identify misconceptions that some/most or all children in my class have.'

'It gave me an insight into how children's ideas can change when given thinking time and how they are able to reason as to why something is right or wrong.'

'It showed concepts in a different format. It got the children to think about their learning.'

The survey findings were corroborated by the feedback we gained from teachers in interviews. For example, teachers said that the Stop and Think game show contestants—animations in the programme—encouraged pupils to reason more which enhanced their learning. Some pupils had told their teachers that using the programme made them think about maths or science. Another view was that Stop and Think helped pupils to further develop social skills such as listening and considering other pupils' points of view. Teachers said that some pupils had taken the Stop and Think idea into other

lessons, that is to say, pupils were taking time to consider questions before answering. Some teachers we interviewed averred that it was difficult to say how using Stop and Think had impacted on their pupils given the limited time they had used the programme and, as one teacher commented, the small size of the pupil cohort (five pupils) in her school.

Views on roll out

Research question: Is the roll out of the intervention feasible for schools?

The survey asked teachers whether they thought Stop and Think in its current form was suitable for roll out to other schools. While 14 gave a positive response, 24 said no and 23 were not sure. Those teachers who gave a 'no' or 'not sure' response were asked to select from a list of possible explanations. The main explanations selected were 'difficult to fit into school day' (23 teachers), 'software problems' (22 teachers), 'pupil engagement' (19 teachers), 'accuracy of content' (18 teachers), 'quality of the animation' (17 teachers) and 'content too easy' (16 teachers).

When we interviewed teachers, they said that the programme was not feasible to use unless it was linked better to the teaching cycle and pupils' learning activities. The topics covered in the sessions were pre-set by the programme. Teachers advocated that the programme should offer more teacher control and more flexibility so that teachers could use it to cover topics the class had covered recently. For example, one teacher suggested that a database of questions would be useful giving teachers the option of tailoring the use of the programme to topics being covered in class. The random nature of the programme makes it difficult to use and to make relevant to what pupils are learning, remarked one teacher. He explained this point by saying that a question on magnets or a questions on animals and bones might pop up just before a maths lesson to which these questions had no direct relevance. He suggested that if the programme has a select button, teachers could use this to align topics with specific learning activities the pupils are currently covering in class.

Theory of Change (TOC)

Research question: Was the theory of change model identified in the pilot an accurate representation of the intervention and its outcomes?

Drawing on all the evidence collected in this evaluation, we conclude that the TOC (Appendix E) devised in the pilot was not an entirely accurate representation of the intervention and its outcomes. The TOC identified the purpose of the intervention as: *'To provide schools with an educational resource that helps to improve pupils' ability to deal with counterintuitive concepts and improve their reasoning in maths and science'*. The survey of teachers, case studies, and follow-up teacher interviews confirmed that the intervention was delivered to the target groups—primary school pupils in Year 3 and Year 5—using a computer-based programme as specified in the TOC. The evaluation found that the output—a reliable, computer-based learning tool—was not wholly achieved as many teachers experienced logistical problems in using the programme. As a result, they could not always use the programme productively in maths and science lessons, which was specified as an intended outcome from the intervention.

The TOC identified the programme's impact as generating *'improvements in pupils' counterintuitive learning reasoning skills and increased attainment in maths and science at Key Stage 2'*. The findings from the trial suggest that the Stop and Think intervention did not make any statistically significant difference in pupils' attainment in maths. In contrast, the intervention demonstrated a positive impact on pupils' attainment in science. Looking at the analysis for each year group separately, the intervention showed that there was no significant effect on Year 3 pupils' maths or science attainment or on Year 5 pupils' maths attainment. Conversely, the intervention contributed positively to science attainment of Year 5 pupils when compared with pupils from combined control groups. The evaluation assessed whether the Stop and Think intervention, which focused on inhibiting maths and science intuitive reasoning, would lead to improvements in inhibitory control in another domain, as measured using the

Chimeric Animal Stroop task. This was not found to be the case as no difference in Chimeric Animal Stroop task outcomes were found between groups.

Control group activity

The control-plus group delivering See+

The findings from the survey and interviews with teachers delivering See+ are presented below.

Fidelity

The survey explored fidelity in the delivery of the See+ programme. A majority (23) of the 32 teachers who responded reported that their class received See+ for ten weeks and four more indicated that their class did not receive See+ for ten weeks but received it sometimes within the ten-week delivery period. Half of the teachers (16) reported that their class received See+ three times a week and 14 reported that this happened sometimes. A majority of the teachers surveyed (24) indicated that their class did not receive See+ at the start of a maths or science lesson, as requested by the researchers. Our observations of See+ sessions in three case-study schools revealed that the sessions were run at different times of the school day and not during the maths and science lessons such as at the end of a lesson, after morning assembly, and following a lunch break. Seventeen of the 32 teachers delivering See+ who responded to the survey question 'To what extent did you discuss Stop and Think (the other Unlocked programme running in your school) with the class?' said that they discussed Stop and Think with their class.

In the two schools where we observed See+ sessions as part of the case studies, they were whole-class sessions. In one school, the teacher read out the question and pupils voted (with hands up) on the answer. The teacher announced the majority decision on the answer and entered this on the computer. In the other school, the teacher would ask one pupil what he/she thought the answer was and then would ask the other pupils in the class whether they agreed or not. The teacher would also use a show of hands, running through the answers and enter the majority vote on the computer. In facilitating the sessions, the teachers did not lead discussion but in one case the teacher explained why the answer given was correct or not. The pupils were engaged, though, in one mixed Year 3 and Year 4 class: the Year 3 pupils who were being targeted were engaged while Year 4 pupils sat quietly during the session.

Suitability

Suitability for the curriculum

While half (16) of the teachers considered that the See+ content (subject matter) was appropriately aligned with the personal, health and social education (PHSE) curriculum, ten were not sure and six said it was not aligned.

Suitability for pupils in the year groups

The survey asked teachers whether, in their view, See+ was suitable for pupils in upper Key Stage 1 (Year 2). A majority of teachers thought the programme was suitable: eight said 'to a great extent', 12 'to some extent', and five 'to a little extent'. Only one teacher said 'not at all'. There were similar results from a question asking teachers whether See+ was suitable for pupils in lower Key Stage 2 (Years 3 and 4): five indicated 'to a great extent', 17 'to some extent', and four 'to a little extent'. In contrast, fewer teachers considered that See+ was suitable for pupils in upper key Stage 2 (Years 5 and 6): two said 'to a great extent', five 'to some extent', and 13 'to a little extent'. Nine teachers did not consider that See+ was suitable for these pupils.

Suitability for pupils with SEN

Most of the teachers surveyed thought that See+ was suitable for pupils with SEN: seven said 'to a great extent', 14 'to some extent', and three 'to a little extent'.

Suitability of questions and graphics

When we interviewed teachers, some said that See+ programme was too simple for Year 5 pupils.

They also pointed out that the quality of the programme's graphics was not very good and should be more age-appropriate to maintain pupils' attention and keep them engaged. Pupils sometimes found it difficult to decipher the characters' facial expressions in the animations which made it hard to understand the scenarios.

Software issues

The survey investigated whether there had been any problems with the software (for example, slow to load and screen freezing) during the delivery of See+ sessions. While half of the teachers surveyed (16) indicated that they had seldom experienced problems seldom (ten teachers) or never (six teachers), half said that they experienced problems often (eight teachers), sometimes (seven teachers), or always (one teacher). Where teachers had experienced problems with the software, they were asked whether this impacted on their pupils' ability to engage fully with the See+ programme. All thought that the software problems had impacted negatively on pupils' engagement.

In two of the three See+ sessions we observed as part of the case studies, there were practical issues. In one case, the programme would not load despite an off-line version being installed in the school. The teacher said that owing to regular loading issues, the three Year 5 classes (up to 60) often did the See+ session at the same time. The group of pupils would then be split into smaller groups to discuss answers to questions in the See+ programme. The teacher acknowledged that this was not ideal as the group was too large but took this action to make it work. In the other case, the pupils found it difficult to read the characters in the programme's animation. The teacher was unsure whether all the questions had been answered when the session had finished because the programme ended abruptly.

Training and resources

The survey asked for teachers' views on the training and resources associated with the programme. When asked whether the training provided by Birkbeck staff was suitable for preparing them to deliver See+, the majority (26) of the teachers considered that the training was suitable: eight said 'highly suitable' and 18 said 'suitable'. Only three teachers said the training was 'not very suitable'. (Three teachers had not received the training). While the survey did not ask for details of the See+ training, our interviews with teachers indicated that training mainly focused on the technical set-up and running of the programme.

When asked how good the written Teacher Guide was in supporting them to deliver See+, over half gave a positive response, five said 'very good' and 12 said 'good', while ten considered it 'acceptable'. A majority of the teachers surveyed (21) indicated that their school did not require any additional resources to run See+, ten teachers were not sure, and one confirmed that the school required software resources.

Perceptions of impact on pupils and teachers

Pupils

The survey asked teachers whether See+ had impacted on their pupils, inviting them to select from a list of possible impacts. A majority (18 teachers) selected developing PHSE skills, followed by taking time to consider their response before answering questions (12 teachers), improving engagement in

learning (two teachers) and enhancing confidence (one teacher). Other impacts identified by teachers (open question) included 'discussing scenarios and looking at and from different perspectives', 'thinking about the way others feel', and 'some increase in participation from less confident children'.

Teachers

The survey asked teachers to what extent they think See+ had a positive impact on them as a class teacher. Just over half (18) indicated a positive impact: eight said 'to some extent' and ten said 'to a little extent'. Fourteen teachers indicated that the programmes had not had a positive impact on them. Examples of positive impact expressed by teachers were:

'See+ enabled me to encourage the children to think about others and make considerations of those around them.'

'It has made me think about ensuring that children develop their empathy skills by being exposed to examples of various social situations.'

'It has shown me social situations and how my class would react or their attitude towards them. This then informs me of how to help deal with playground or other issues that may come up.'

Views on roll-out

A majority of the teachers surveyed (22) did not think that See+ in its current form was suitable for roll-out to other schools, two teachers thought it was suitable and eight were not sure. Those teachers who gave a 'no' or 'not sure' response were asked to select from a list of possible explanations. The main explanations selected were: 'quality of animation' (29), 'content too easy' (15 teachers), 'difficult to fit into school day' (15 teachers), 'pupil engagement' (13 teachers), 'accuracy of the content' (12 teachers), and 'software problems' (11 teachers). When interviewed, teachers said that See+ would not be suitable for roll-out unless the quality of the animation was improved.

Conclusion

Interpretation

This trial had two primary outcomes which increases the risk that a false positive result may be found through chance. The maths outcome did not reach statistical significance and the intervention demonstrated a positive impact on pupils' attainment in science: the intervention group pupils scored higher on the GL Assessment PTS when compared to the combined control group pupils (control and control-plus groups). This evidence comes from the combined science analysis.

Looking at the analysis for each year group separately, the intervention showed no statistically significant effect on Year 3 pupils' maths or science attainment; this was true when the intervention pupils were compared to the combined control group pupils and pupils from the control-plus group only. The intervention also did not show any significant effect on Year 5 pupils' maths attainment when comparing to the combined control groups, but did show a significant improvement compared to the control-plus group only. The intervention contributed positively to the science attainment of Year 5 pupils when compared with the pupils from the combined control group as well as when compared with the control-plus group pupils only. Improvement in science attainment in Year 5 was also associated with the number of Stop and Think sessions performed.

Analysis from the Stroop assessment revealed no evidence that the intervention made a difference in pupils' general inhibitory control function. Moreover, the CACE findings also indicated that insufficient implementation was not the underlying cause of null results.

There were mixed results for pupils who had been eligible to receive free school meals (FSM) any point in the previous six years. For Year 3 and Year 5 maths, and Year 5 science, FSM pupils made additional progress, on average, compared to the control group. This was not the case for Year 3 science pupils; they made no more additional progress than the control group. However, the study was not powered to measure an effect for FSM pupils and the effects were not significant.

Additional analyses that looked at differences between Stop and Think and See+ showed that Stop and Think had a positive effect on combined maths and combined science attainments of the intervention group compared to the See+ group. The main purpose of having a control-plus group using See+ was to examine whether improvement in pupil attainment was just a result of using a novel computer-based programme at the start of maths and science lessons rather than specifically due to Stop and Think. This was an active control and suggests that the Stop and Think intervention had specific effects on maths and science achievement in Year 5, beyond any effects that may stem from the children and teachers knowing they were taking part in an intervention and working on a computerised intervention together.

The mixed findings suggest that the intervention did not wholly achieve the intended impact specified in the Theory of Change. The evaluation also found that the output—a reliable computer-based learning tool—was not wholly achieved as many teachers experienced logistical problems in using the programme. As a result, they could not always use the programme productively in maths and science lessons, which was specified as an intended outcome from the intervention. Findings from this evaluation suggest that while the teachers thought that the Stop and Think software aligned well with the maths and science curricula, pupils found that the science questions were appropriately challenging whereas the maths questions were quite easy. There was no option for teachers to choose topics and/or subject in the software and therefore when pupils were faced with repeated questions, which were relatively easy, they lost interest. However, feedback from the schools suggested that the characters from the intervention software encouraged pupils to reason more than they would normally, which

enhanced their learning and made them think. The challenging questions kept their motivation and interest in the 'game show' whilst helping them learn how to learn new topics in science. Another possible reason for non-significant results in maths could be the length of the intervention. It could be entirely possible that it takes longer than ten weeks to encourage pupils to inhibit common misconceptions in maths and to learn counterintuitive concepts. It is unlikely that the non-significant results in maths are due to the training or implementation of the intervention as the schools were content with the level of training and support they received. The majority of schools delivered the intervention as intended with the fidelity to intervention being moderate to high in most cases. We can also rule out the teacher effect in one subject over the other as pupils within a class were randomised to take either maths test or a science test. Results from the CACE analysis support the evidence shown by the main analyses. Number of Stop and Think sessions was not associated with Year 3 pupils' maths or science attainment or Year 5 pupils' maths attainment. However, higher number of sessions was associated with greater impact on Year 5 pupils' science attainment as demonstrated in the main analysis of this outcome.

Limitations

Out of 87 schools, primary outcomes data was collected from 84 schools. The primary ITT analyses included 84% of trial pupils. Only a small proportion of this attrition was due to school dropout, and the majority of pupils who were lost to follow-up were lost due to reasons unrelated to the intervention. These included pupils leaving the school before testing, pupils being absent on the day of testing, or not being able to match pupils to NPD data (to obtain prior attainment). When we regressed whether the pupil was missing at follow-up or not, a number of covariates were significantly associated with the outcome. The pattern of missing data demonstrated that the data was not missing completely at random (MCAR) and therefore we undertook multiple imputation. Results from the imputed models showed similar results to that of the substantive model. Therefore, we are fairly confident in the results presented in the ITT models.

While fidelity to intervention implementation was moderate to high in most cases, schools reported a number of limitations. Schools described several issues with the software, which impeded smooth delivery of the intervention. Teachers often mentioned that the quality of the animation was poor which made it difficult for them to facilitate the session and engage pupils. Some of the questions in the software were pitched too low for pupils of this age and the repetitive nature of the questions made it difficult to retain pupil interest. Some teachers found it difficult to fit Stop and Think sessions in their busy timetables. When prompted, the majority of teachers did not endorse the roll-out of the programme to other schools owing to difficulty in fitting delivery into the school day, software problems, pupil engagement, the accuracy of content, quality of animation, and some of the content being too easy.

Future research and publications

If future work is considered on the computer programme, we recommend that the content of the software is also revised keeping in mind the age group of the intended recipients. The software will require a number of changes in order to make the programme a more productive and engaging experience for teachers and pupils. The Stop and Think software would benefit from having more advanced animations and graphics to engage pupils. The Stop and Think programme could also be improved if it was linked better to the teaching cycle and pupils' learning activities. This would mean that the programme will need to be more flexible and be able to offer teachers greater control over choosing the topics, subject area, and level of difficulty.

References

- *Allen, M. (2014) *Misconceptions in Primary Science*, United Kingdom: McGraw-Hill Education.
- Angrist, J. D. and Imbens, G. W. (1995) 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of the American Statistical Association*, 90 (430), pp. 431–442. DOI 10.1080/01621459.1995.10476535
- *Babai, R., Shalev, E. and Stavy, R. (2015) 'A warning intervention improves students' ability to overcome intuitive interference', *ZDM*, 47 (5), pp. 735–745. DOI 10.1007/511858-015-0670-y
- *Bofferding L. (2019) 'Understanding negative numbers', in Norton A., Alibali M. (eds), *Constructing Number: Research in Mathematics Education*. DOI 10.1007/978-3-030-00491-0_12
- Borenstein, M., Hedges, L. V., Higgins, J. P. T. and Rothstein, H. R. (2009) *Introduction to Meta-Analysis*, Chichester: John Wiley and Sons.
- *Borst, G., Simon, G., Vidal, J., and Houdé, O. (2013) 'Inhibitory control and visuospatial reversibility in Piaget's seminal number conservation task: a high-density ERP study.' *Frontiers in Human Neuroscience*, 7, 920 [online]. DOI 10.3389/fnhum.2013.00920.
- *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 [online]. DOI 10.1111/cogs.12126
- *Brault Foisy, L-M., Potvin, P., Riopel, M. and Masson, S. (2015) 'Is inhibition involved in overcoming a common physics misconception in mechanics?', *Trends in Neuroscience and Education*, 4 (1–2), pp. 26–36. DOI 10.1016/j.tine.2015.03.001
- Confederation of British Industry (2017) '“Helping The UK Thrive”: CBI/Pearson Education and Skills Survey 2017'. <https://cbicdnend.azureedge.net/media/1171/cbi-educating-for-the-modern-world.pdf?v=20190905.2>
- *Diamond, A. and Lee, K. (2011) 'Interventions shown to aid executive function development in children 4 to 12 years old', *Science*, 333, pp. 959–964. DOI 10.1126/science.1204529
- *Diamond, A. and Ling, D. S. (2016) 'Conclusions about interventions, programs, and approaches for improving executive functions that appear justified and those that, despite much hype, do not', *Developmental Cognitive Neuroscience*, 18, pp. 34–48. DOI 10.1016/j.dcn.2015.11.005.
- *Dunbar, K. N., Fugelsang, J. A. and C. S. (2007) 'Do naive theories ever go away? Using brain and behavior to understand changes in concepts', in Lovett, M. C. and Shah, P. (eds), *Thinking with Data: 33rd Carnegie Symposium on Cognition*, Hillsdale, NJ: Erlbaum.
- Education Endowment Foundation (2013) 'Pre-testing in EEF Evaluations'. https://educationendowmentfoundation.org.uk/public/files/Evaluation/Writing_a_Protocol/Pre-testing_paper.pdf [31 January, 2018].
- Education Endowment Foundation (2018) 'Statistical Analysis Guidance for EEF Evaluations'. https://educationendowmentfoundation.org.uk/public/files/Grantee_guide_and_EEF_policies/Evaluation/Writing_a_Protocol_or_SAP/EEF_statistical_analysis_guidance_2018.pdf
- *Evans, J. S. B. T. (2003) 'In two minds: dual-process accounts of reasoning,' *Trends in Cognitive Sciences*, 7, pp. 454–459. DOI 10.1016/j.tics.2003.08.012

- Fischer Family Trust (no date) 'FFT Development Paper: KS1 Estimates Based on EYFSP (09/11)'.
<http://csapps.norfolk.gov.uk/cssshared/ecourier2/fileoutput.asp?id=11608>
- *Fugelsang, J. A. and Dunbar, K. N. (2005) 'Brain-based mechanisms underlying complex causal thinking', *Neuropsychologia*, 43, pp. 1204–213. DOI 10.1016/j.neuropsychologia.2004.10.012
- *Goel, V., and Dolan, R. J. (2003) 'Explaining modulation of reasoning by belief'. *Cognition*, 87, pp. B11–22.
- HM Government: Department for Business, Energy and Industrial Strategy (2017) *Industrial Strategy: Building a Britain Fit for the Future*, Industrial Strategy White Paper, CM 9528.
<https://www.gov.uk/government/publications/industrial-strategy-building-a-britain-fit-for-the-future>
- Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S. and Gernsbacher, M. A. (2007) 'The science of sex differences in science and mathematics', *Psychological Science in the Public Interest*, 8 (1), pp. 1–51. DOI 10.1111/j.1529-1006.2007.00032.x
- House of Commons, Business, Innovation and Skills Committee (2015) 'The Government's Productivity Plan. Second Report of Session (2015-16)', HC 466.
<https://publications.parliament.uk/pa/cm201516/cmselect/cmbis/466/466.pdf>
- *Houdé, O. and Tzourio-Mazoyer, N. (2003) 'Neural foundations of logical and mathematical cognition', *Nature Reviews Neuroscience*, 4, pp. 507–514.
- *Houdé, O., Zago, L., Mellet, E., Moutier, S., Pineau, A., Mazoyer, B. and Tzourio-Mazoyer, N. (2000) 'Shifting from the perceptual brain to the logical brain: the neural impact of cognitive inhibition training', *Journal of Cognitive Neuroscience*, 12, pp. 721–728.
- Hutchison, J. E., Lyons, I. M. and Ansari, D. (2019) 'More similar than different: Gender differences in children's basic numerical skills are the exception not the rule', *Child Development*, 90 (1), pp. 66–79. DOI 10.1111/cdev.13044
- *Kerns, A. K., Eso, K. and Thomson, J. (1999) 'Investigation of a direct intervention for improving attention in young children with ADHD', *Developmental Neuropsychology*, 16, pp. 273–295.
- 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>
- *Kusché, C. A. and Greenberg, M. T (1994) *The PATHS Curriculum*, Seattle: Developmental Research and Programs.
- *Linzarini, A., Houdé, O. and Borst, G. (2015) 'When Stroop helps Piaget: An intertask positive priming paradigm in 9-Year-old children', *Journal of Experimental Child Psychology*, 139, pp. 71–82. DOI 10.1016/j.jecp.2015.05.010
- *Lubin, A., Vidal, J., Lanoë, C., Houdé, O. and Borst, G. (2013) 'Inhibitory control is needed for the resolution of arithmetic word problems: A developmental negative priming study.' *Journal of Educational Psychology*, 105, 3, 701–708 [online]. DOI 10.1037/a0032625.
- Mareschal, D. (2016) 'The neuroscience of conceptual learning in science and mathematics', *Current Opinion in Behavioural Sciences*, 10, pp. 14–18. DOI 10.1016/j.cobeha.2016.06.001

- *Masson, S., Potvin, P., Riopel, M. and Brault Foisy, L-M. (2014) 'Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity', *Mind, Brain, and Education*, 8 (1), pp. 44–55. DOI 10.1111/mbe.12043
- *McClelland, J. L. and Rogers, T. T. (2003) 'The parallel distributed processing approach to semantic cognition', *Nature Reviews Neuroscience*, 4 (4), pp. 310–322.
<http://www.cnbc.cmu.edu/~plaut/IntroPDP/papers/McClellandRogers03NatNeuRev.semCog.pdf>
- McNamara, S., Roy, P. and Rutt, S. (2018) 'Statistical Analysis Plan for the Evaluation of Learning Counterintuitive Concepts'.
https://educationendowmentfoundation.org.uk/public/files/Projects/Neuroscience_-_Counterintuitive_Concepts_SAP.pdf
- National Foundation for Educational Research (2016) 'Protocol for the Evaluation of Counterintuitive Concepts Intervention'.
https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_Learning_counterintuitive_concepts_Final.pdf
- National Foundation for Educational Research (2018) 'Protocol for the Evaluation of Counterintuitive Concepts Intervention'.
https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Neuroscience_-_Counterintuitive_Concepts_Protocol_AMENDED.pdf
- National Foundation for Educational Research (no date) 'Privacy Notice for the Evaluation of Counterintuitive Concepts Intervention (UnLocke)'.
https://www.nfer.ac.uk/media/3258/eec_privacy_notice.pdf
- Organisation for Economic Co-operation and Development (2018) 'PISA 2015: PISA results in focus'.
<https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf>
- *O'Reilly, R. C., Herd, S. A. and Pauli, W. M. (2010) 'Computational models of cognitive control', *Current Opinion in Neurobiology*, 20 (2), pp. 257–261. DOI 10.1016/j.conb.2010.01.008
- Parker, P. D., Van Zanden, B. and Parker, R. B. (2018) 'Girls get smart, boys get smug: Historical changes in gender differences in math, literacy, and academic social comparison and achievement', *Learning and Instruction*, 54, pp. 125–137. DOI 10.1016/j.learninstruc.2017.09.002
- Penner, A. M. and Paret, M. (2008) 'Gender differences in mathematics achievement: Exploring the early grades and the extremes', *Social Science Research*, 37 (1), pp. 239–253. DOI 10.1016/j.ssresearch.2007.06.012.
- *Prado, J. and Noveck, I. A. (2007) 'Overcoming perceptual features in logical reasoning: A parametric functional magnetic resonance imaging study', *Journal of Cognitive Neuroscience*, 19, pp. 642–657.
- *Riggs, N. R., Greenberg, M. T., Kusché, C. A. and Pentz, M. A. (2006) 'The mediational role of neurocognition in the behavioral outcomes of a social-emotional prevention program in elementary school students: effects of the PATHS Curriculum', *Prevention Science*, 7, pp. 91–102.
- *Rousselle, L., Palmers, E., and Noël, M. P. (2004) 'Magnitude comparison in preschoolers: What counts? influence of perceptual variables', *Journal of Experimental Child Psychology*, 8 (1), pp. 57–84. DOI 10.1016/j.conb.2010.01.008

- *Shipstead, Z., Hicks, K. L. and Engle, R. W. (2012) 'Cogmed working memory training: Does the evidence support the claims?', *Journal of Applied Research in Memory and Cognition*, 1, pp. 185–193. DOI 10.1016/j.jarmac.2012.06.003
- Spierer, L., Chavan, C. and Manuel, A. L. (2013) 'Training-induced behavioral and brain plasticity in inhibitory control', *Frontiers in Human Neuroscience*, 7, p. 427. DOI 10.3389/fnhum.2013.00427.
- *Stavy, R., and Babai, R. (2010) 'Overcoming intuitive interference in mathematics: Insights from behavioral, brain imaging and intervention studies', *ZDM*, 42 (6), pp. 621–633.
- *Stavy, R. and Tirosh, D. (2000) *How Students (Mis-)understand Science and Mathematics*, New York: Teachers College Press.
- *Thorell, L. B., Lindqvist, S., Bergman Nutley, S., Bohlin, G. and Klingberg, T. (2009) 'Training and transfer effects of executive functions in preschool children', *Developmental Science*, 12 (1), pp. 106–113. DOI 10.1111/j.1467-7687.2008.00745.x
- UnLocke (2016) 'Learning counterintuitive concepts. An innovative maths and science learning activity informed by neuroscience'. <http://unlocke.org/neuroscience.html>
- *Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A. and Kyrianakis, G. (2018a) 'The recruitment of shifting and inhibition in on-line science and mathematics tasks', *Cognitive Science*, 42, pp. 1860–1886. DOI 10.1111/cogs.12624.
- *Vosniadou, S., Pnevmatikos, D. and Makris, N. (2018b) 'The role of executive function in the construction and employment of scientific and mathematical concepts that require conceptual change learning', *Neuroeducation*, 5 (2), pp. 62–72. DOI 10.24046/neuroed.20180502.62
- *Wass, S., Porayska-Pomsta, K. and Johnson, M. H. (2011) 'Training attentional control in infancy', *Current Biology*, 21, pp. 1543–1547. DOI 10.1016/j.cub.2011.08.004
- Wright, I., Waterman, M., Prescott, H. and Murdoch-Eaton, D. (2003) 'A new Stroop-like measure of inhibitory function development: typical developmental trends', *Journal of Child Psychology and Psychiatry*, 44 (4), pp. 561–575.

* Please note that these references have been provided by neuroscience experts at Birkbeck College.

Appendix A: EEF cost rating

Cost ratings are based on the approximate cost per pupil per Year of implementing the intervention over three Years. More information about the EEF's approach to cost evaluation can be found [here](#). Cost ratings are awarded as follows:

Cost rating	Description
£ £ £ £ £	<i>Very low:</i> less than £80 per pupil per Year.
£ £ £ £ £	<i>Low:</i> up to about £200 per pupil per Year.
£ £ £ £ £	<i>Moderate:</i> up to about £700 per pupil per Year.
£ £ £ £ £	<i>High:</i> up to £1,200 per pupil per Year.
£ £ £ £ £	<i>Very high:</i> over £1,200 per pupil per Year.

Appendix B: Security classification of trial findings

Learning Counterintuitive Concepts - Maths

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	Power	Attrition*					
5	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%					
4	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%	4		Adjustment for Balance [N/A]		4
3	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%					
2	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%			Adjustment for threats to internal validity [N/A]		
1	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%					
0	No comparator	MDES > 0.6	>50%					

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 4 padlocks (good experimental design, appropriate analysis; MDES at randomisation: 0.14; attrition from pupils randomised to pupils analysed: 17%, resulting in loss of 1 padlock)
- **Reason for adjustment for balance** (if made): N/A
- **Reason for adjustment for threats to validity** (N/A)
- **Final padlock score:** initial score adjusted for balance and internal validity = 4 padlocks

*Attrition should be measured at the pupil level, even for cluster trials and from the point of randomisation to the point of analysis.

Learning Counterintuitive Concepts - Science

Rating	Criteria for rating			Initial score		Adjust		Final score
	Design	Power	Attrition*					
5	Well conducted experimental design with appropriate analysis	MDES < 0.2	0-10%					
4	Fair and clear quasi-experimental design for comparison (e.g. RDD) with appropriate analysis, or experimental design with minor concerns about validity	MDES < 0.3	11-20%	4		Adjustment for Balance [N/A]		4
3	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%					
2	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%			Adjustment for threats to internal validity [N/A]		
1	Comparison group with poor or no matching (E.g. volunteer versus others)	MDES < 0.6	41-50%					
0	No comparator	MDES > 0.6	>50%					

- **Initial padlock score:** lowest of the three ratings for design, power and attrition = 4 padlocks (good experimental design, appropriate analysis; MDES at randomisation: 0.14; attrition from pupils randomised to pupils analysed: 16%, resulting in loss of 1 padlock)
- **Reason for adjustment for balance** (if made): N/A
- **Reason for adjustment for threats to validity** (N/A)
- **Final padlock score:** initial score adjusted for balance and internal validity = 4 padlocks

*Attrition should be measured at the pupil level, even for cluster trials and from the point of randomisation to the point of analysis.

Appendix C: Main trial information and consent



1st September, 2017

Dear Parent or Guardian

We are writing to let you know that your child's school has agreed to take part in the national evaluation of a novel computer-based maths and science learning activity. This project is funded by the *Wellcome Trust* and the *Education Endowment Foundation (EEF)*, and is being led by Prof Denis Mareschal from *Birkbeck University of London*. Through the use of a computerised game, this learning activity aims to help children "stop and think" when faced with new maths and science ideas, rather than just going straight to what they might initially believe to be true when answering questions. This moment's hesitation can help them take on the new information, and thus learn new concepts more effectively. As part of their normal lessons after half term in October 2017 until February 2018, pupils in your child's class will be asked to either complete some maths and science activities or some social skills learning activities on a computer.

The *EEF* have appointed the *National Foundation for Educational Research (NFER)* to evaluate how effective the UnLocke learning activity is in improving children's maths and science academic outcomes. The *NFER* is a leading independent provider of educational research. They will work closely with the UnLocke researchers to collect maths and science achievement data on all the children who take part in this study. The primary purpose of these data is to evaluate the UnLocke project. The achievement data will also be made available to the schools. Any results from the study will remain anonymous. The school will provide your child's name, date-of-birth and unique pupil identifier to the Birkbeck team who will pass this on to *NFER* via a secure portal. This information will be treated in strictest confidence by the *NFER* and Birkbeck.

Where will my child's data be sent?

All of the data collected in this project, including data about your child that is retrieved from the National Pupil Database (e.g., gender, free school meal status... but not their name), will be provided to Fischer Family Trust (FFT, which is the organisation appointed to manage *EEF*'s data archive) and stored in the *EEF* data archive and the UK Data Archive for research purposes. The overall findings will be included, with no reference to the school's or individual children's names, in a publicly available report used to influence practice nationally.

This project is confidential and no school or child will be named in any report of this work.

This letter is to ensure that you are happy for your child to take part in these evaluations. We do hope that you will agree to allow your child to take part in this project. Volunteer participation is essential for any kind of educational research to progress. We can have the best ideas in the world, but without participants to test the ideas, nothing can move forward.

Henry Wellcome Building
Birkbeck, University of London
Malet Street
London
WC1E 7HX
unlocke@psychology.bbk.ac.uk





IF YOU DO NOT WISH YOUR CHILD TO PARTICIPATE, PLEASE SIGN BELOW AND RETURN THIS FORM TO YOUR SCHOOL BY THE END OF THE WEEK.

If you are happy for your child to participate in this evaluation, you **do not need to do anything**.

Please feel free to get in touch if you have any further questions regarding the study. Additional details of the project, including the science behind the project and who is involved, can be found at www.unlocke.org. Please direct any further queries to unlocke@psychology.bbk.ac.uk

Further details of the NFER can be found at www.nfer.ac.uk/

Further details of the EEF can be found at <https://educationendowmentfoundation.org.uk/>

Many thanks for your time and attention.

Best regards,

THE UNLOCKE TEAM

I do not wish my child to participate in this study on inhibition control in science and maths.

Child's name: _____ Child's Class: _____

Parent's name: _____

Parent's signature: _____

Date: _____

Henry Wellcome Building
Birkbeck, University of London
Malet Street
London
WC1E 7HX
unlocke@psychology.bbk.ac.uk



Privacy notice for the evaluation of counterintuitive concepts intervention (UnLocke)

1. Why are we collecting this data?

Personal data is being collected to enable the evaluation of the 'Stop and Think' and 'SEE+' element of 'UnLocke' using a randomised controlled trial. The main aim of 'Stop and Think' programme is to improve learner's ability to adapt to counterintuitive concepts via training the learner to inhibit their initial response and instead, give a more delayed and reflective answer to ultimately improve learners' educational outcomes. The trial aims to ascertain the impact of the intervention on pupil attainment in maths and science.

2. Who is this research project sponsored and funded by?

The Education Endowment Foundation (EEF) and the Wellcome Trust commissioned Birkbeck College to develop and deliver UnLocke in collaboration with UCL Institute of Education.

National Foundation for Educational Research (NFER) is undertaking the independent evaluation which is funded by EEF. NFER and Birkbeck College are the joint data controller for this evaluation.

3. What is the legal basis for processing activities?

The legal basis for processing personal data is covered by:

GDPR Article 6 (1) (f) which states that 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

Our legitimate interest for processing your personal data is to administer the randomised control trial.

4. How will personal data be obtained?

Birkbeck College is responsible for recruiting schools for this trial. They will collect teacher and pupil personal data from the participating schools. They will share this data with NFER using NFER's secure data exchange portal.

NFER will obtain background pupil data from the Department for Education's (DfE) National Pupil Database (NPD) using DfE's secure data exchange portal.

NFER will undertake case study visits in up to six schools where we will interview individual staff members involved in the project and observe Stop and Think and SEE+ sessions.

NFER will also administer an online teacher questionnaire via *Questback*.

NFER and Birkbeck will assess the pupils at the end of the trial using the GL assessment tests and inhibitory function development tests administered in the schools.

5. What personal data is being collected by this project and how it will be shared between the research partners?

Personal data for the main trial will include data about teachers and pupils from the participating schools as described below:

Teacher data: Birkbeck College will collect data (name, job title and contact details) about a nominated lead teacher and class teachers in Years 3 and 5 in each participating school so that NFER can liaise with the individual about the evaluation. NFER will administer an online teacher questionnaire where names and email addresses will be collected to monitor the response rate. Birkbeck or EEF will not see any data from the teacher questionnaire. No personal data from the case study visits will be collected.

Birkbeck will collect personal data about pupils. This includes pupil names, date of births and UPNs. This will be shared with NFER in order for us to access pupil background data held by the DfE's NPD. The NPD data that we will request covers pupil prior attainment at the end of Key Stage 1, pupil free school meal eligibility and gender. Birkbeck will not see any data from the NPD.

NFER will administer GL Assessment tests. Each participating pupil will only take one test—either the Progress test in Maths (PTM) or the Progress tests in Science (PTS). GL-Assessment, acting as a data processor, mark the tests. NFER will share the test results with Birkbeck College.

Birkbeck College's Research Assistants will assess the pupils on their inhibitory function development. Once the test administration is complete, they will share the test results with NFER.

NFER will match all of the above pupil data to pupil assessment data. The assessment data includes pupil results from the GL Assessment tests and pupil results from the inhibitory function development. Above datasets will enable NFER to undertake primary and secondary outcomes analyses in order to achieve the aims mentioned in section 1.

NFER will share all of the above pupil data (pupil names, dates of birth, UPN matched to the NPD data described above and assessment results) with EEF's data archive partner—Fischer Family Trust. Anonymised data will also be stored in the UK Data Archive.

6. Is personal data being transferred outside of the European Economic Area (EEA)?

No personal data is stored or transferred outside of the EEA.

7. How long will personal data be retained?

NFER and Birkbeck College will delete any personal data after three Years from completion of the project. (Note that retention of personal data is subject to agreement by the NPD team at DfE).

NFER will send all the data to FFT archive within three months of the end of the project who will keep the data, and take responsibility for data protection compliance.

8. Can I stop my personal data being used?

NFER handles your personal data in accordance with the rights given to individuals under data protection legislation. If at any time you wish us to withdraw your data or correct errors in it, please contact Tom Dickinson at **unlocketrial@nfer.ac.uk**.

In certain circumstances, data subjects have the right to restrict or object processing, please contact our **Compliance Officer** at **compliance@nfer.ac.uk**. They also have the right to see information held about them. NFER will cooperate fully when a subject access request (SAR) is made.

9. Who can I contact about this project?

NFER and Birkbeck are responsible for the day-to-day management of this project. Contact Tom Dickinson at **unlocketrial@nfer.ac.uk** or Professor Denis Mareschal at **d.mareschal@bbk.ac.uk** with any queries.

In certain circumstances, data subjects have the right to restrict or to object to data processing, please contact NFER's **Compliance Officer** at **compliance@nfer.ac.uk** or Birkbeck College's Data Protection Officer David McElroy at **d.mcelroy@bbk.ac.uk**. They also have the right to see information held about them. You can make a subject access request by contacting either organisation.

If you have a concern about the way this project processes personal data, we request that you raise your concern with us in the first instance (see the details above). Alternatively, you can contact the Information Commissioner's Office, the body responsible for enforcing data protection legislation in the UK, at **<https://ico.org.uk/concerns/>**.

Appendix D: Memorandum of Understanding

UnLocke Randomised Controlled Trial Study

MEMORANDUM OF UNDERSTANDING

Aims of the evaluation

The aim of this project is to evaluate the impact of a computer-based learning activity called UnLocke on Year 3 and Year 5 children's attainment in maths and science. One hundred schools in England will participate in this project. This research is funded by the EEF (Education Endowment Foundation) and the Wellcome Trust. The results of the research will contribute to our collective understanding of the potential value of the learning activity and will be made publicly available.

The project

Using a randomised controlled trial (RCT) design, Year 3 and Year 5 classes from participating schools across the country will be randomly selected to be in an 'intervention group' or a 'control' group for a ten week programme. The intervention group will use the UnLocke learning activity at the start of science and maths lessons. There will be two control groups in this project so that the impact of the learning activity can be evaluated and compared. The first control group will use another computerised learning activity (SEE+) during PSHE lessons focusing on social skills. The second control group will continue their maths and science lessons as normal - a "business as usual" approach. In summary, some classes will therefore be randomly allocated to the maths and science learning activities, some to the social skills learning activity and some to carrying on as usual. In a single school one year group will be in the intervention group and the other year group will be in one of the two control groups.

During this project, you will be contacted by both the BirkbeckProject Team, who are responsible for training and supporting participating teachers and supplying the materials to schools, and by the NFER (National Foundation for Education Research), who are carrying out an independent evaluation of the project.

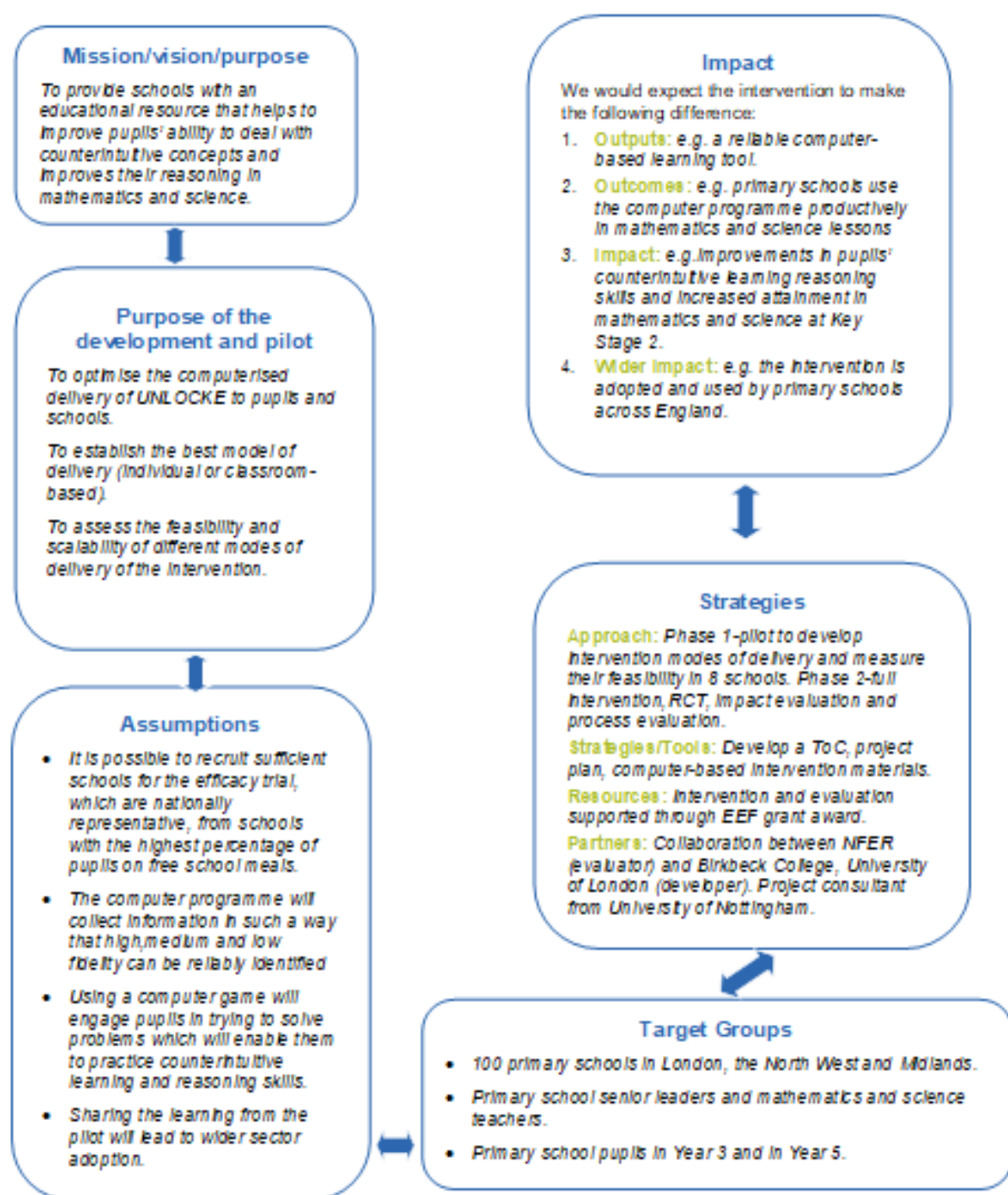
This memorandum of understanding (MoU) explains what your school's participation in the study will entail. If you agree to take part and accept the terms and conditions outlined, please sign a copy of this form and either scan and return by email or return it by postal mail to the contact provided at the end of this letter using the FREEPOST envelope provided for your convenience.

Structure of the evaluation

Schools will be involved in delivering two approaches, with all Year 3 and Year 5 classes in your school being randomly assigned to one of the three approaches for the whole 10 week programme:

1. *UnLocke Science and Maths learning activity:* Classes from one year group (Year 3 or Year 5) will use the UnLocke science and maths learning activity **at the beginning of their maths and/or science lessons for 15 minutes, three times a week, for 10 weeks.**
2. Classes from the other year group (Year 3 or Year 5) will act as an important control group and will be randomly allocated by the NFER to either:
3. *Social Skills learning activity:* Classes in this group will use a similar computerised learning activity designed to improve social skills (SEE+), **during PSHE lessons, or other appropriate times during the school day, but not during maths and science lessons, for 15 minutes, three times a week, for 10 weeks.**

Appendix E: Theory of Change (ToC)* for the External Evaluation of Learning Counterintuitive Concepts



* 'A ToC shows a path from needs to activities to outcomes to impact. It describes the change you want to make and the steps involved in making that change happen. A ToC helps organisations to begin to measure their impact, ultimately getting more out of their resources to help more people.' Kolk, A. and Lumley, T. (2012). *Theory of change: The beginning of making a difference*. London: New Philanthropy Capital.

Appendix F: Randomisation code (SPSS syntax)

* Encoding: windows-1252.

* Encoding: .

dataset close all.

*Code for EECC Randomisation.

*CODE REPEATS 3 TIMES IN THIS SCRIPT.

*1x Schools w 1 form entry.

*1x Schools w 2 form entry.

*1x Schools w All Other form entry.

*Read the data file in.

GET

FILE='I:\EECC\Data from Birkbeck\Randomisation\First
Wave\Clean_Combined_Post_Randomisation.sav'.

sort cases by dfe Year.

list dfe totalclass.

*Above list generates the same values as the one that was sent to Birkbeck (below file).

*Use that file instead.

dataset close all.

GET DATA

/TYPE=XLSX

/FILE='I:\EECC\Data from Birkbeck\Randomisation\First
Wave\NFER_Clean_First_Randomisation_For_Birkbeck_Confirmation.xlsx'

```
/SHEET=name 'School & teacher data'
```

```
/CELLRANGE=FULL
```

```
/READNAMES=ON
```

```
/DATATYPEMIN PERCENTAGE=95.0.
```

```
EXECUTE.
```

freq dfe Year.

```
aggregate outfile=* mode=ADDVARIABLES /break dfe /
```

```
totalclass=n.
```

```
aggregate outfile=*/break dfe Year/
```

```
SCHOOLNAME=first(SCHOOLNAME)/
```

```
URN=first(URN)/
```

```
CONTACT=first(CONTACT)/
```

```
ROLE=first(ROLE)/
```

```
EMAIL=first(EMAIL)/
```

```
PHONE=first(PHONE)/
```

```
FORMENTRY=first(FORMENTRY)/
```

```
MIXEDYEARS=first(MIXEDYEARS)/
```

```
totalclass=first(totalclass).
```

```
list dfe totalclass.
```

*If the school has less than one class, the school is not eligible to take part in the trial, remove them.

```
cross Year by totalclass.
```

temp.

select if totalclass=1.

freq dfe urn.

select if dfe<>'305/3916'.

exe.

save outfile='i:/temp/eeccall.sav'.

```
*****
*****
.

*****
*****
.

*****
*****
.
```

get file='i:/temp/eeccall.sav'.

***First randomise one form schools.

cross totalclass by FORMENTRY by Year.

*Remove schools where there are only Year 3 and not Year 5.

select if FORMENTRY='1' and totalclass=2.

freq YEAR.

freq FORMENTRY.

cross FORMENTRY by Year.

*Use the Year group aggregate code for Form2/3 and Other.

match files file=*/first=f/last=l by DFE YEAR.

cross f by l.

select if f= 1.

NOTE

*USE APPROP CODE FROM LIST BELOW, Comment out others.

dataset copy Form1.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

*Generate school level random variable.

aggregate outfile=*/break=DFE/nYears=n(YEAR).

list var=DFE nYears.

*Change mtindex for each randomisation.

set rng=mt, mtindex=172017.

compute schrand=rv.uniform(0,1).

*list var=schrnd.

****NOTE.

*CHANGE TO APPROPRIATE NAME.

match files file=Form1/table=*/in=inschrnd by dfe.

freq inschrnd.

dataset close all.

*Need to randomise Year groups such that half of Year 3 receives intervention and the other half in control group.

*However, this needs to be random so use schrand variable.

sort cases by schrand.

match files file= */first=FX/last=LX by schrand.

cross FX by LX.

aggregate outfile* mode=addvariables/

totcase=n.

*For even number of cases (which this will be as there are Year 3 and 5 both present in the dataset), run the following code where the less than first half of Y3 gets int.

*If we used FX=1 and \$casenum le (totcase/2) y3int=1, the Y5 in same school gets Int too which is not intended.

if FX=1 and \$casenum lt (totcase/2) y3int=1.

if FX=1 and \$casenum ge (totcase*0.75) y3int=2.

if LX=1 and \$casenum gt (totcase/2) y5int=1.

if LX=1 and \$casenum le (totcase/4) y5int=2.

recode y3int (sysmiss=0).

recode y5int (sysmiss=0).

freq var y3int y5int.

if sum (y3int, y5int)=1 group=1.

if sum (y3int, y5int)=2 group= 2.

if sum (y3int, y5int)=0 group= 3.

add value labels group 1 'Int' 2'Control' 3'See+'.

*Check for 2:1:1.

freq group.

cross YEAR by FORMENTRY.

cross YEAR by group.

cross FORMENTRY by group.

sort cases by DFE YEAR.

*Save to Correct File.

save outfile='i:/temp/eecc1form.sav'.

dataset close all.

```
*****
*****
*****
*****
*****
*****
```

2 FORM ENTRY CODE.

get file='i:/temp/eeccall.sav'.

cross totalclass by FORMENTRY by Year.

select if FORMENTRY='2' and totalclass=4.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

sort cases by DFE YEAR.

*Use the Year group aggregate code for Form2/3 and Other.

match files file=*/first=f/last=l by DFE YEAR.

cross f by l.

select if f= 1.

dataset copy Form2.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

*Need to randomise Years groups to intervention and control.

*Then randomise within control to C and See+.

*Below is an attempt to randomised Int and Control with code from stratified2.

*First randomise schools so Intervention will not happen for both Years of one school.

aggregate outfile=*/break=DFE/nYears=n(YEAR).

list var=DFE nYears.

*Change mtindex for each randomisation.

set rng=mt, mtindex=2017100.

compute schrand=rv.uniform(0,1).

*list var=schrnd.

dataset name sch_rand.

****NOTE.

*CHANGE TO APPROPRIATE NAME.

match files file=Form2/table=*/in=inschrnd by dfe.

freq inschrnd.

dataset close all.

*Need to randomise Year groups such that half of Year 3 receives intervention and the other half in control group.

*However, this needs to be random so use schrand variable.

sort cases by schrand.

match files file=*/first=FX/last=LX by schrand.

cross FX by LX.

aggregate outfile* mode=addvariables/

totcase=n.

*For even number of cases (which this will be as there are Year 3 and 5 both present in the dataset), run the following code where the less than first half of Y3 gets int.

*If we used FX=1 and \$casenum le (totcase/2) y3int=1, the Y5 in same school gets Int too which is not intended.

if FX=1 and \$casenum lt (totcase/2) y3int=1.

if FX=1 and \$casenum ge (totcase*0.75) y3int=2.

if LX=1 and \$casenum gt (totcase/2) y5int=1.

if LX=1 and \$casenum le (totcase/4) y5int=2.

recode y3int (sysmiss=0).

recode y5int (sysmiss=0).

freq y3int y5int.

if sum (y3int, y5int)=1 group=1.

if sum (y3int, y5int)=2 group= 2.

if sum (y3int, y5int)=0 group= 3.

add value labels group 1 'Int' 2 'Control' 3 'See+'.

*Check for 2:1:1.

freq group.

freq Year.

freq formentry.

cross YEAR by FORMENTRY.

cross YEAR by group.

cross FORMENTRY by group.

sort cases by DFE YEAR.

*Save to Correct File.

save outfile='i:/temp/eccc2form.sav'.

dataset close all.

```
*****
*****
*****
*****
*****
*****
```

***Three form and 6 classes.

get file='i:/temp/eccall.sav'.

cross totalclass by FORMENTRY by Year.

select if FORMENTRY='3' and totalclass=6.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

sort cases by DFE YEAR.

*Use the Year group aggregate code for Form2/3 and Other.

match files file=*/first=f/last=l by DFE YEAR.

cross f by l.

select if f= 1.

dataset copy Form3.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

*Need to randomise Years groups to intervention and control.

*Then randomise within control to C and See+.

*Below is an attempt to randomised Int and Control with code from stratified2.

*First randomise schools so Intervention will not happen for both Years of one school.

aggregate outfile=*/break=DFE/nYears=n(YEAR).

list var=DFE nYears.

*Change mtindex for each randomisation.

set rng=mt, mtindex=1002016.

compute schrand=rv.uniform(0,1).

*list var=schrnd.

dataset name sch_rand.

****NOTE.

*CHANGE TO APPROPRIATE NAME.

match files file=Form3/table=*/in=inschrand by dfe.

freq inschrand.

dataset close all.

*Need to randomise Year groups such that half of Year 3 receives intervention and the other half in control group.

*However, this needs to be random so use schrand variable.

sort cases by schrand.

match files file=*/first=FX/last=LX by schrand.

cross FX by LX.

aggregate outfile* mode=addvariables/

totcase=n.

*For even number of cases (which this will be as there are Year 3 and 5 both present in the dataset), run the following code where the less than first half of Y3 gets int.

*If we used $FX=1$ and $\$casenum \leq (totcase/2)$ $y3int=1$, the Y5 in same school gets Int too which is not intended.

if $FX=1$ and $\$casenum \leq (totcase/2)$ $y3int=1$.

if $FX=1$ and $\$casenum \geq (totcase*0.75)$ $y3int=2$.

if $LX=1$ and $\$casenum \geq (totcase/2)$ $y5int=1$.

if $LX=1$ and $\$casenum \leq (totcase/4)$ $y5int=2$.

recode $y3int$ (sysmiss=0).

recode $y5int$ (sysmiss=0).

freq $y3int$ $y5int$.

if sum (y3int, y5int)=1 group=1.

if sum (y3int, y5int)=2 group= 2.

if sum (y3int, y5int)=0 group= 3.

add value labels group 1 'Int' 2'Control' 3'See+'.

*Check for 2:1:1.

freq group.

freq Year.

freq formentry.

cross YEAR by FORMENTRY.

cross YEAR by group.

cross FORMENTRY by group.

sort cases by DFE YEAR.

*Save to Correct File.

save outfile='i:/temp/eecc3form.sav'.

dataset close all.

```
*****
*****
*****
*****
*****
*****
```

***ALL OTHER ENTRY CODE.

get file='i:/temp/eeccall.sav'.

cross totalclass by FORMENTRY by Year.

select if (FORMENTRY='1' and totalclass<>2)

or (FORMENTRY='2' and totalclass<>4)

or (FORMENTRY='3' and totalclass<>6).

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

sort cases by DFE YEAR.

*Use the Year group aggregate code for Form2/3 and Other.

match files file= */first=f/last=l by DFE YEAR.

cross f by l.

select if f= 1.

dataset copy Other.

freq YEAR.

freq FORMENTRY.

cross YEAR by FORMENTRY.

*Need to randomise Years groups to intervention and control.

*Then randomise within control to C and See+.

*Below is an attempt to randomised Int and Control with code from stratified2.

*First randomise schools so Intervention will not happen for both Years of one school.

```
aggregate outfile=*/break=DFE/nYears=n(YEAR).
```

```
list var=DFE nYears.
```

```
*Change mtindex for each randomisation.
```

```
set rng=mt, mtindex=20173100.
```

```
compute schrand=rv.uniform(0,1).
```

```
*list var=schrand.
```

```
dataset name sch_rand.
```

```
match files file=Other/table=*/in=inschrand by dfe.
```

```
freq inschrand.
```

```
dataset close all.
```

*Need to randomise Year groups such that half of Year 3 receives intervention and the other half in control group.

*However, this needs to be random so use schrand variable.

```
sort cases by schrand.
```

```
match files file=*/first=FX/last=LX by schrand.
```

```
cross FX by LX.
```

```
aggregate outfile* mode=addvariables/
```

```
totcase=n.
```

*For even number of cases (which this will be as there are Year 3 and 5 both present in the dataset), run the following code where the less than first half of Y3 gets int.

*If we used $FX=1$ and $\$casenum \leq (totcase/2)$ $y3int=1$, the Y5 in same school gets Int too which is not intended.

```
if FX=1 and $casenum lt (totcase/2) y3int=1.
```

```
if FX=1 and $casenum ge (totcase*0.75) y3int=2.
```

if LX=1 and \$casenum gt (totcase/2) y5int=1.

if LX=1 and \$casenum le (totcase/4) y5int=2.

recode y3int (sysmiss=0).

recode y5int (sysmiss=0).

****NOTE****.

*CODE BELOW ONLY NEEDED FOR CLEAN OTHER DATA SUBSET.

*This Prevent double coding of single Y3 schools.

*if sum (FX, LX)=2 y5int= 0.

freq y3int y5int.

if sum (y3int, y5int)=1 group=1.

if sum (y3int, y5int)=2 group= 2.

if sum (y3int, y5int)=0 group= 3.

freq group.

add value labels group 1 'Int' 2'Control' 3'See+'.

*Check for 2:1:1.

freq group.

freq Year.

freq formentry.

cross YEAR by FORMENTRY.

cross YEAR by group by dfe.

cross FORMENTRY by group by Year.

sort cases by DFE YEAR.

*Save to Correct File.

save outfile='i:/temp/eeccallother.sav'.

dataset close all.

**Add these files together.

add files

FILE='i:/temp/eecc1form.sav'/in=in1form/

/FILE='i:/temp/eecc2form.sav'/in=in2form/

/FILE='i:/temp/eecc3form.sav'/in=in3form/

/FILE='i:/temp/eeccallother.sav'/in=inother.

freq in1form in2form in3form inother.

***Match with the original file to double check we are not missing out on any dfe and Year groups.

sort cases by dfe Year.

match files file=*/file='i:/temp/eeccall.sav'/in=inorig by dfe Year/map.

freq inorig.

save outfile='K:\EECC\cfs\randomisation\First wave\first wave_r.sav'/keep dfe to totalclass nYears group.

***Checking.

get file='K:\EECC\cfs\randomisation\First wave\first wave_r.sav'.

cross YEAR by group.

cross FORMENTRY by group.

sort cases by dfe Year.

list dfe Year group.

cross Year by group by dfe.

***Further checking.

aggregate outfile=*/break=dfe/

group1=first(group)/

group2=last(group)/

ntot=n.

freq dfe ntot.

list dfe group1 group2.

compute wrong=0.

if group1=group2 wrong=1.

freq wrong.

***No schools have same group assignment for both Year groups.

*Ready file to upload on the portal.

get file='K:\EECC\cfs\randomisation\First wave\first wave_r.sav'.

SAVE TRANSLATE OUTFILE='K:\EECC\cfs\randomisation\First wave\first wave_r.xlsx'

/TYPE=XLS

/VERSION=12

```
/MAP
```

```
/FIELDNAMES VALUE=NAMES
```

```
/CELLS=LABELS
```

```
/REPLACE.
```

```
output save outfile='K:\EECC\cfs\randomisation\First wave\first wave_r.spv'.
```

Appendix G: Stop and Think and SEE+ teacher survey

Qa/	
0.	<p>Evaluation of the learning counterintuitive concepts ('Unlocke') intervention: Stop and Think / SEE+</p> <p>The Education Endowment Foundation (EEF) has commissioned NFER to undertake a survey of teachers participating in the Unlocke learning counterintuitive concepts project which is being led by Birkbeck College. The purpose of this survey is to explore how the Stop and Think / SEE+ programme has been implemented and whether it has met its aims. The survey will inform our overall assessment of the impact of the Stop and Think programme; your views are invaluable to us so please take the time to complete this survey.</p> <p>All responses will be treated in confidence and reported only in aggregated or anonymised form. The information collected will be used for research purposes only and will not be shared with EEF or Birkbeck College.</p> <p>This survey will take five to ten minutes to complete.</p> <p>If you have any queries, please contact NFER on 01753 XXXXXX or unlocketrial@nfer.ac.uk</p>

Qi – (ASK ALL)				
i	Did you deliver:	Please select one box only.	1	Stop and Think?
			2	SEE+?
			3	Neither Stop and Think nor SEE+.

Qii – (ASK ALL)				
ii	Do you teach	Please select one box only.	1	Year 2/3?
			2	Year 3?
			3	Year 3 /4?
			4	Year 4/5?
			5	Year 5?
			6	Year 5/6
			7	Other (please specify)

If selected Qi 3 (Neither) Please send respondent to the SUBMIT PAGE. They should only complete Qi and Qii.

Delivering Stop and Think / SEE+ in class

Q1 – (ASK IF Qi = 1 (S&T); 2 (See+))			
1	On the whole, who delivered the Stop and Think / SEE+ sessions to your class?	Please select one box only.	1 Teacher
			2 Teaching assistant
			3 Someone else (Please specify)

Q2 – (ASK ALL)			
2	Did the same person deliver all Stop and Think / SEE+ sessions?	Please select one box only.	1 Yes
			2 Not sure
			3 No

Q3 – (ASK ALL)					
Did your class receive Stop and Think / SEE+:					
Q3		Please select one box per row.			
		A Yes	B Sometimes	C No	D Not sure
3.1	for ten weeks?				
3.2	three times a week?				
3.3	at the start of a maths / science lesson?				

Q4 – (ASK STOP AND THINK only)			
In your view, was the content (subject matter) of Stop and Think appropriate for your class for:			
Q4		Please select one box per row.	
		A Maths	B Science
4.1	Yes, it was pitched appropriately		
4.2	No, it was too easy		
4.3	No, it was too difficult		
4.4	Not sure		

Q5a (STOP AND THINK)

Is the Stop and Think **content** (subject matter) appropriately aligned with the curriculum for:

Q5		Please select one box per row.		
		A Yes	B Not sure	C No
5.1	Maths			
5.2	Science			

Q5b (SEE+ – ASK IF SEE+ SCHOOL)

5	Is the SEE+ content (subject matter) appropriately aligned with the PHSE curriculum?	Please select one box only.	1	Yes
			2	Not sure
			3	No

Q6 – (ASK ALL)

6	In your view, is Stop and Think / SEE+ suitable for:	Please select one box on each row.				
		[1] A great extent	[2] To some extent	[3] To a little extent	[4] Not at all	[5] Not sure
6.1	pupils in upper KS1 (Year 2)?					
6.2	pupils in lower KS2 (Years 3 and 4)?					
6.3	pupils in upper KS2 (Years 5 and 6)?					
6.4	pupils with SEN?					

Software

Q7– (ASK ALL)

7	How often, if at all, did you experience problems with the software (e.g. slow to load, screen freezing, other) during the Stop and Think / SEE+ sessions?	Please select one box only.				
1	Always					
2	Often					

3	Sometimes			
4	Seldom			
5	Never [go to Q9]			
8– ASK IF Q7 = 1 (Almost always), Q7 = 2 (Often), Q7 = 3 (Sometimes), Q7 = 4 (Seldom)				
8	If you experienced issues with the software, to what extent do you think this impacted on the pupils' ability to engage fully with the Stop and Think / SEE+ programme?	Please select one box only.	1	A great extent
2			To some extent	
3			To a little extent	
4			Not at all	
5			I did not experience issues with the software	

Training and resources

Q9– (ASK ALL)				
9	In your view, was the training provided by Birkbeck staff suitable in preparing you to deliver Stop and Think / SEE+?	Please select one box only.	1	Highly suitable
2			Suitable	
3			Not very suitable	
4			Not at all suitable	
5			Did not receive training [Go to Q11]	

Q10 –(Ask IF Q9 = 3 (Not very suitable), Q9 = 4 (Not at all suitable))		
10	Do you have suggestions for improvements to the training?	(Please write your response in the box below.)

Q11a – (ASK ALL)				
11a	In your view, how good was the written Teacher Guide in supporting you to deliver Stop and Think/ SEE+?	Please select one box only.	1	Very good
2			Good	
3			Acceptable	
4			Poor	
5			Very poor	
6			Did not use	

Q11b – (OR, Ask IF Q11 = 4 (Poor), Q11 = 5 (Very poor))		
11b	You indicated the Teacher Guide was poor / very poor in supporting you to deliver Stop and Think / SEE+.	(Please explain your answer in the box below.)

12a – (ASK ALL)				
12a	Did your school require any additional resource/s to run Stop and Think / SEE+?	Please select one box only.	1	Yes
			2	Not sure
			3	No

Q12b – (MR, Ask IF Q12 = 1 (Yes))				
12b	What additional resource/s did your school require?	(Please select all that apply)	12b.1	Hardware (e.g. computers, white board etc)
			12b.2	Software (e.g. Firewall updates)
			12b.3	Staff time
			12b.4	Other (please specify)

Perceptions of impact

Q14– (ASK ALL of stop and think participants)				
14	In your opinion, to what extent did Stop and Think have a positive impact on your class's maths ability?	Please select one box only.	1	A great extent
			2	To some extent
			3	To a little extent
			4	Not at all
			5	Not sure

Q15– (ASK ALL of stop and think participants)				
15	In your opinion, to what extent did Stop and Think have a positive impact on your class's science ability?	Please select one box only.	1	A great extent
			2	To some extent
			3	To a little extent
			4	Not at all
			5	Not sure

Q16a (STOP & THINK)				
16	What other impact/s, if any, did Stop and Think have on pupils in your class?	(Please select all that apply)	16.1	Taking time to consider their response before answering questions
			16.2	Improving engagement in learning
			16.3	Enhancing confidence
			16.4	Developing numeracy skills
			16.5	Developing science skills
			16.6	Other (please specify)

Q16b (SEE+)

16	What other impact/s, if any, did SEE+ have on pupils in your class?	(Please select all that apply)	16.1	Taking time to consider their response before answering questions
			16.2	Improving engagement in learning
			16.3	Enhancing confidence
			16.4	Developing PHSE skills
			16.5	Other (please specify)

Q17i– (ASK ALL)

17i	In your opinion, to what extent did Stop and Think / SEE+ have a positive impact on you as class teacher?	Please select one box only.	1	A great extent
			2	To some extent
			3	To a little extent
			4	Not at all

Q17ii – (OR, Ask IF Q17i = 1 (A great extent); 2 (To some extent); 3 (To a little extent))

17ii	In the box below, please explain how Stop and Think / SEE+ has had a positive impact on you as a class teacher.	
------	-----------------------------------------------------------------------------------------------------------------	--

Time commitment

Q18 – (ASK all stop and think participants only)									
Question heading									
18	To enable us to provide useful information to other schools about the cost and time involved in delivering Stop and Think in school, please give your best estimate about how much time you spent on the following activities:	Please select one box on each row.							
		[1] Did not do this	[2] 1-5 minutes	[3] 6-15 minutes	[4] 16-20 minutes	[5] 21-30 minutes	[6] 31 – 59 minutes	[7] 1 – 2 hours	[8] Over 2 hours
18.1	Training								
18.2	Preparing for your first Stop and Think session								
18.3	The average time you spent preparing for each Stop and Think session (excluding logging into the session)								

18.4	Setting up a Stop and Think session (including logging in)								
18.5	Delivering a Stop and Think session (once logged in)								
18.6	Other activities involved in Stop and Think. Please list and state time involved								

Suggestions for improvement and roll out

Q19a– (ASK ALL)

19	To what extent did you discuss Stop and Think with colleagues in other Year groups during the ten week intervention?	Please select one box only.	1	A great extent
			2	To some extent
			3	To a little extent
			4	Not at all

Q19b– (ASK ALL)

19	To what extent did you discuss Stop and Think (the other Unlocked programme running in your school) with the class teacher delivering Stop and Think?	Please select one box only.	1	A great extent
			2	To some extent
			3	To a little extent
			4	Not at all

20a – (ASK ALL)

20	Do you think Stop and Think / SEE+ in its current form is suitable for roll out to other schools?	Please select one box only.	1	Yes
			2	Not sure
			3	No

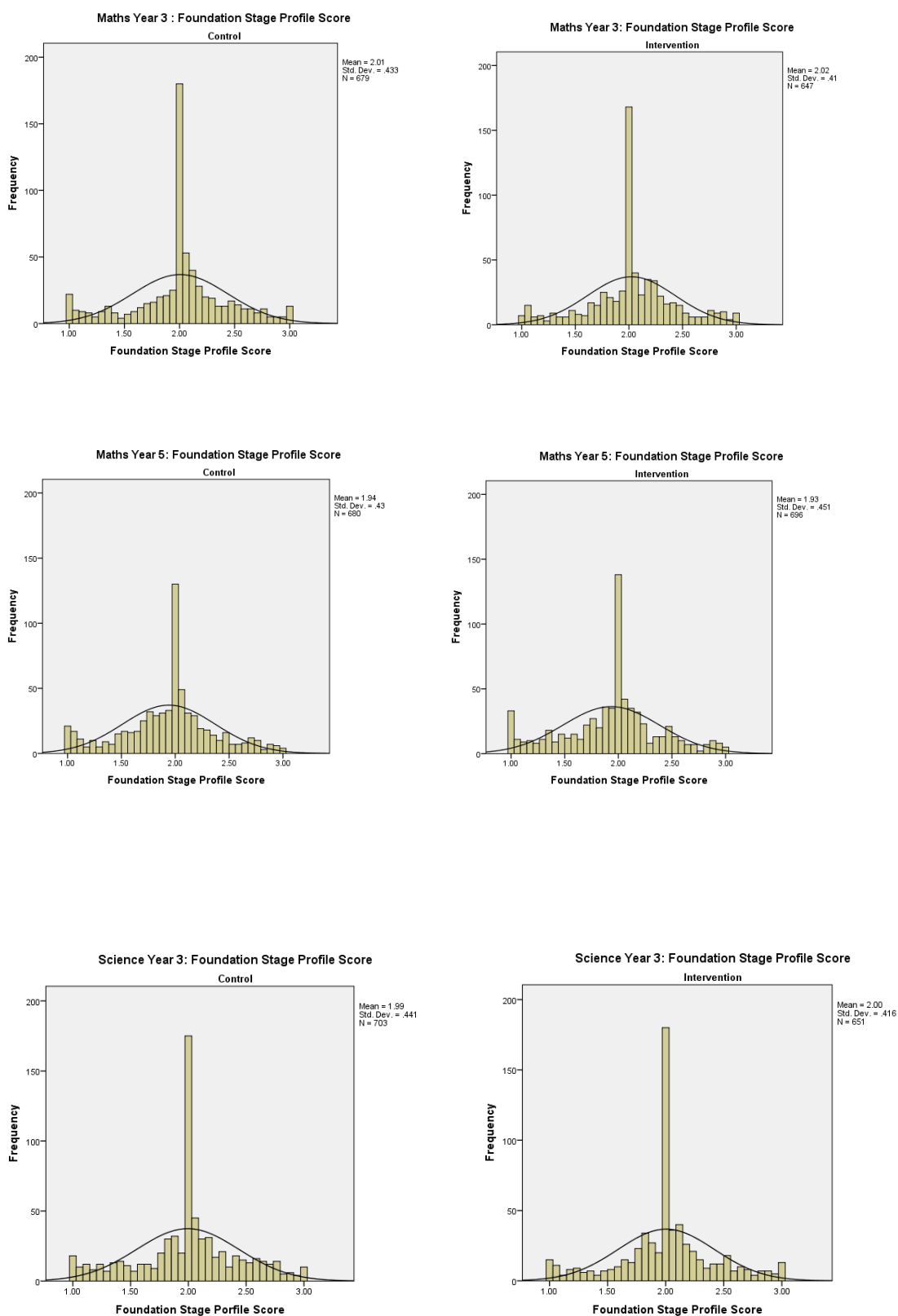
Q20b (MR, ASK IF Q20 = 2 (Not sure), if Q20 = 3 (no))

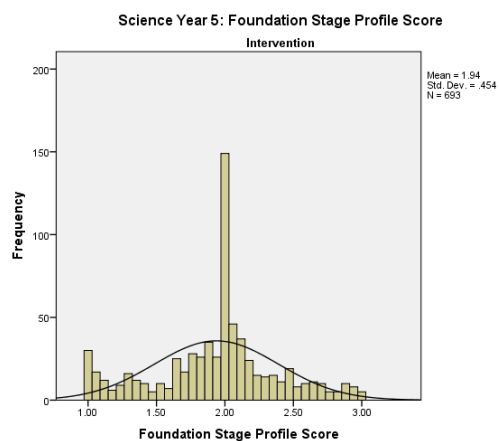
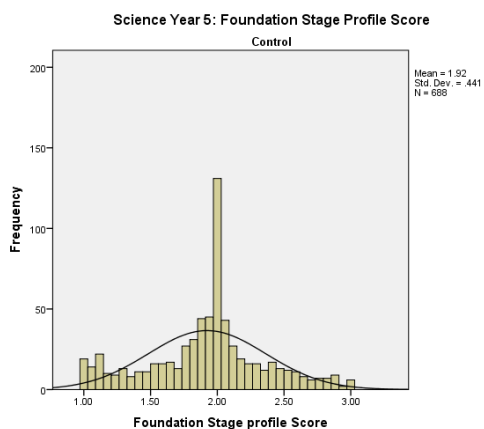
20b	Please explain why not	(Please select all that apply)	20b.1	Quality of the animation
			20b.2	Accuracy of the content
			20b.3	Content too easy
			20b.4	Content too difficult
			20b.5	Software problems (e.g. frozen screen)
			20b.6	Pupil engagement
			20b.7	Difficult to fit into the school day
			20b.8	Other (please specific)

21 – (ASK ALL)				
21	Which version of Stop and Think / SEE+ was running in your class:	Please select one box only.	1	Online
			2	Offline
			3	Not sure

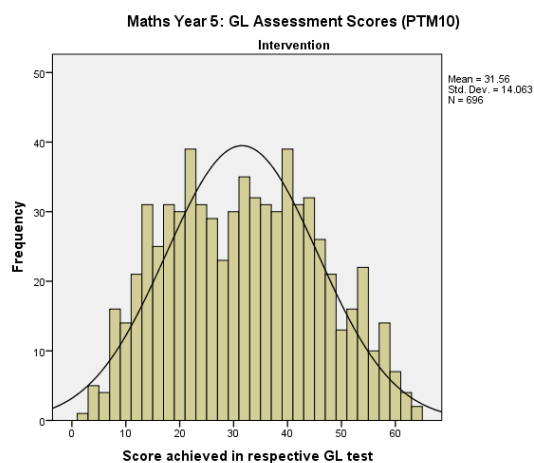
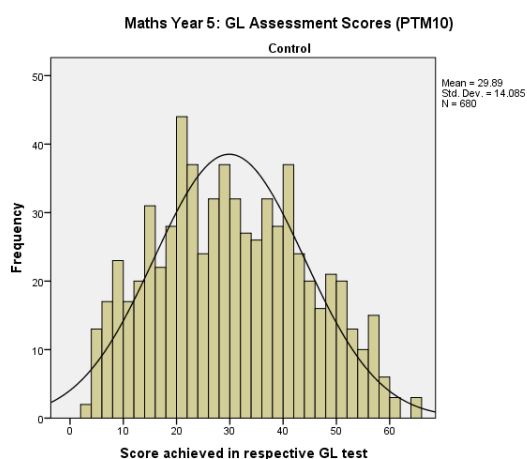
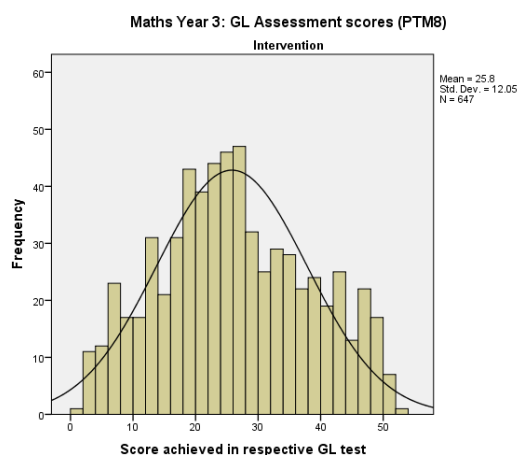
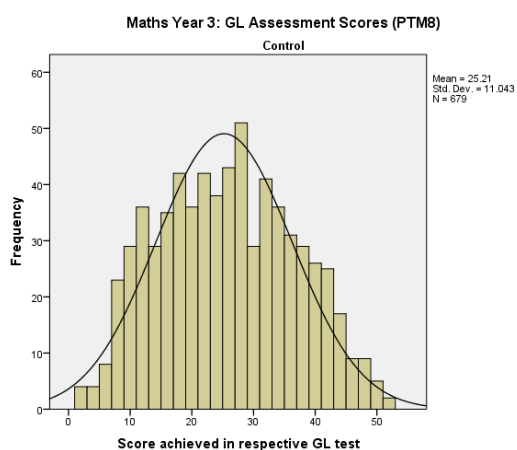
SUBMIT PAGE	
<p>You have reached the end of the survey. Thank you for answering our questions. Please click 'Next' to send your response. Once submitted, you will not be able to go back and change any of your answers.</p>	

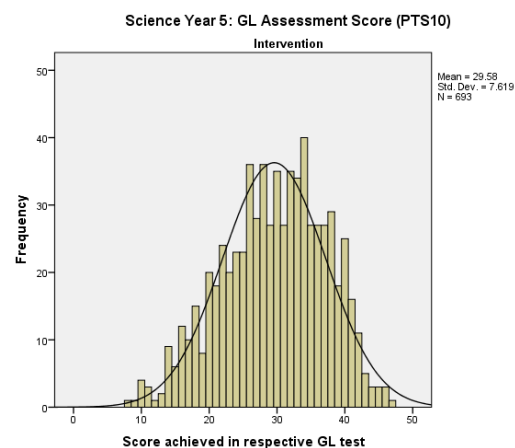
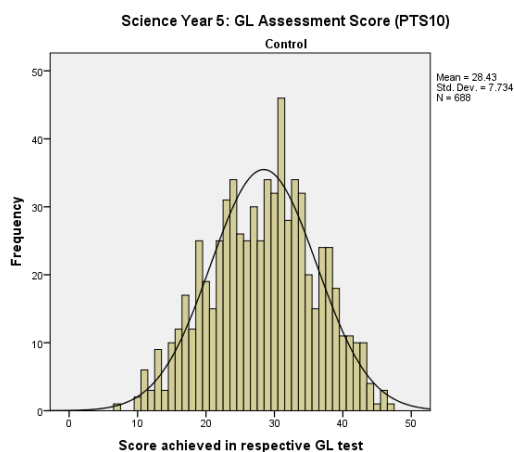
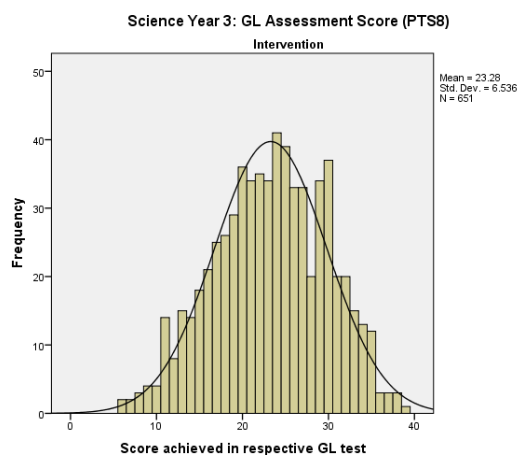
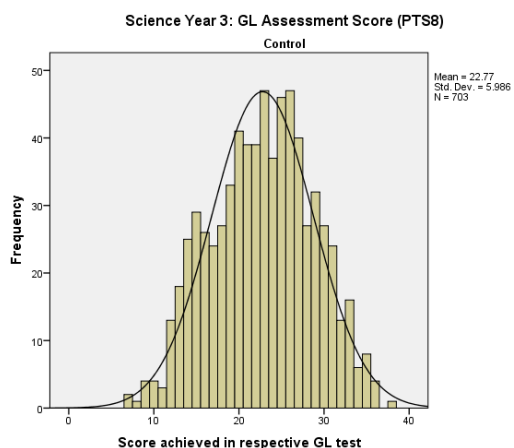
Appendix H: Histograms of Prior Attainment, EYFSP scores



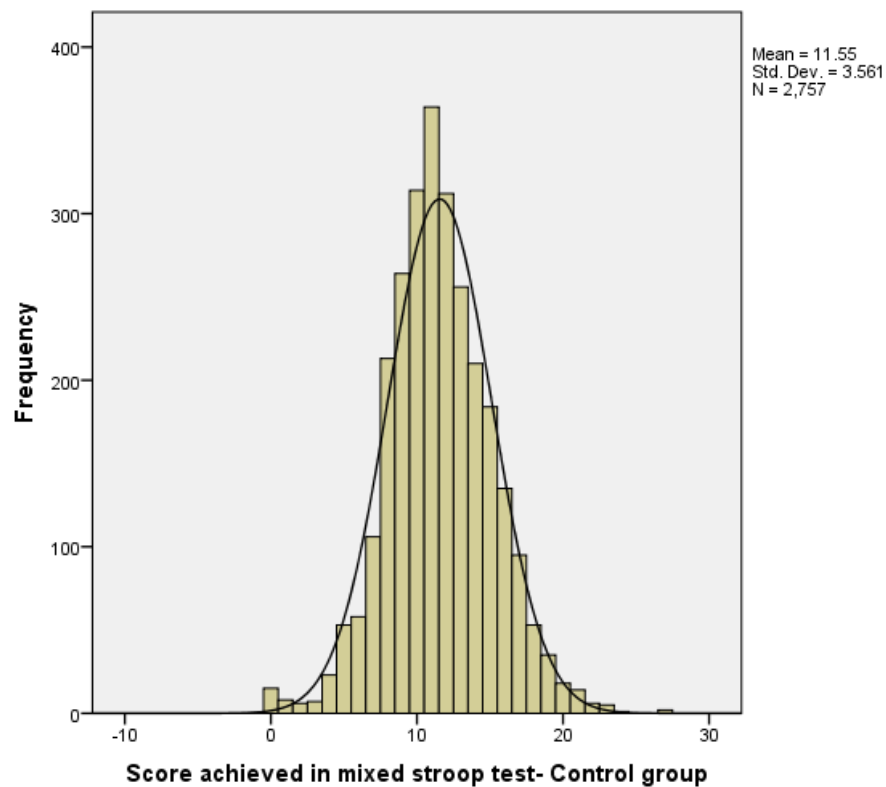


Appendix I: Distribution of outcomes measures (PTM8, PTM10, PTS8 and PTS10) by randomisation groups

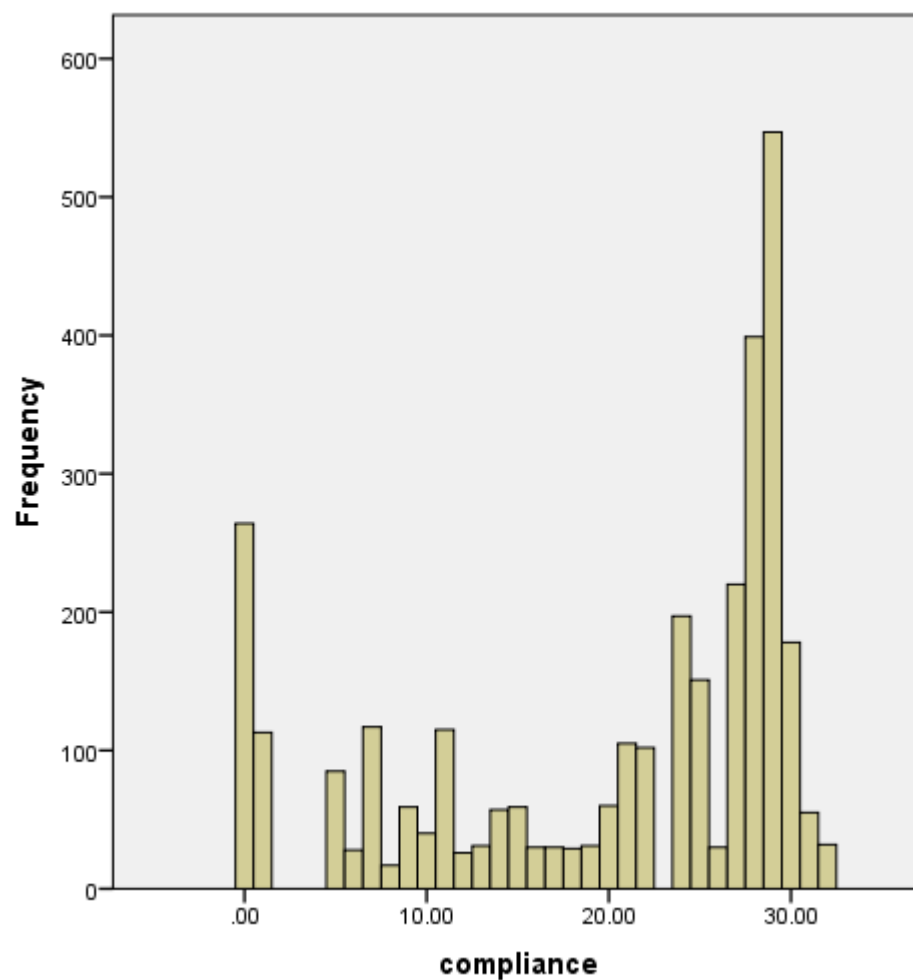




Appendix J: Histograms of secondary outcome measures from the Chimeric Animal Stroop task



Appendix K: Number of Stop and Think sessions (compliance measure)



This work was produced using statistical data from ONS. The use of the ONS statistical data in this work does not imply the endorsement of the ONS in relation to the interpretation or analysis of the statistical data. This work uses research datasets which may not exactly reproduce National Statistics aggregates.

You may re-use this document/publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence v3.0.

To view this licence, visit <https://nationalarchives.gov.uk/doc/open-government-licence/version/3> or email: psi@nationalarchives.gsi.gov.uk

Where we have identified any third-party copyright information you will need to obtain permission from the copyright holders concerned. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

This document is available for download at <https://educationendowmentfoundation.org.uk>



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

<https://educationendowmentfoundation.org.uk>

 [@EducEndowFoundn](https://twitter.com/EducEndowFoundn)

 Facebook.com/EducEndowFoundn