



Education
Endowment
Foundation

ScratchMaths

Evaluation report and executive summary

December 2018

Independent evaluators:

Mark Boylan, Sean Demack, Claire Wolstenholme, John Reidy and Sarah Reaney-Wood

**Sheffield
Hallam
University**

Sheffield
Institute
of Education



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 Trust (now part of Impetus - Private Equity Foundation) 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.



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

Danielle Mason
Head of Research

Education Endowment Foundation
9th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

p: 020 7802 1679

e: danielle.mason@eefoundation.org.uk

w: www.educationendowmentfoundation.org.uk

About the evaluator

The project was independently evaluated by a team from Sheffield Hallam University: Professor Mark Boylan, Sean Demack, Claire Wolstenholme, Dr John Reidy, Sarah Reaney-Wood, Professor Hilary, Povey, Ian Guest and Anna Stevens supported by Martin Culliney and CDARE administrators Chris Roffey of Bebras/Beaver UK provided information on scoring used in the Code Club evaluation.

The lead evaluator was Professor Mark Boylan.

Contact details:

Professor Mark Boylan

Sheffield Institute of Education, Sheffield Hallam University City Campus, Howard Street, Sheffield S1 1WB, UK Email: m.s.boylan@shu.ac.uk

Tel: 0114 225 6130

Contents

Executive summary.....	4
Introduction	6
Methods	17
Impact evaluation	33
Process evaluation.....	48
Conclusions.....	63
References.....	72
Appendix A: EEF cost rating.....	75
Appendix B: Security classification of trial findings.....	76
Appendix C: ScratchMaths content and ScratchMaths team theory of change	77
Appendix D: Consent forms and MoUs.....	79
Appendix E: Detail of team roles	85
Appendix F: Computational thinking test - development and analysis.....	87
Appendix G: Distribution of primary outcome (overall K2 maths attainment, 2017) and follow-on secondary outcomes (attainment in the three KS2 maths papers, 2017)	102
Appendix H: Distribution of interim secondary outcome (Computational thinking test, 2016).....	105
Appendix I: Multilevel analyses & calculation of effect sizes	107
Appendix J: Process evaluation samples and data consolidation.....	110
Appendix K: Fidelity to ScratchMaths & the on-treatment analysis	113
Appendix L: ScratchMaths team post PD evaluation.....	120

Executive summary

The project

ScratchMaths is a two-year computing and mathematics curriculum designed for pupils aged nine to eleven years old, supported by teacher professional development (PD). The programme uses Scratch, a free online programming environment, to integrate coding activities into mathematical learning. Year 5 and 6 teachers or computing teachers received two full days of training in the summer term before using materials the following academic year. In this project, Y5 teachers could also access two optional half-day sessions and Y6 teachers had the opportunity to attend an optional, further half-day of training and an online webinar. Pupils were expected to be taught ScratchMaths for at least one hour every fortnight and were expected to have access to at least one computer for every two pupils to access the online activities. ScratchMaths was developed and delivered by the UCL Knowledge Lab.

This school-level randomised controlled trial measured the computational thinking scores of pupils after one year of the intervention, and Key Stage 2 maths scores after two years. 110 schools were involved at the start of the trial. An implementation and process evaluation consisted of visits to professional development events, telephone interviews, teacher surveys, and review of project data and materials. The trial ran between April 2015 and June 2017.

Key conclusions

1. There is no evidence that ScratchMaths had an impact on pupils' KS2 maths outcomes. This result has a very high security rating.
2. Children in ScratchMaths schools made additional progress in computational thinking scores at the end of Year 5, compared to children in the other schools. The additional progress was higher for children who have ever been eligible for free school meals.
3. Many schools did not fully implement ScratchMaths, particularly in Year 6. High fidelity to the intervention was found in 44% of schools in Y5 and 24% in Y6. Implementation was enhanced where schools provided teachers with time to work through materials.
4. Teachers viewed ScratchMaths as a good way of addressing aspects of the primary computing curriculum, good for improving Scratch programming skills, good professional development, and good for its high quality materials. Five teachers voiced concerns that the lower-attaining pupils needed additional support or adaptation of materials to fully access all ScratchMaths content.
5. Participation in professional development and the use of materials is potentially a very low-cost per pupil option to enhance non-specialists' knowledge and skills to teach aspects of the primary computing curriculum in a manner that is suitable for boys and girls.

EEF security rating

These findings have a very high security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of schools. The trial was a well-designed two-armed randomised controlled trial. The trial was well powered and relatively few (7%) pupils who started the trial were not included in the final analysis. The pupils in ScratchMaths schools were similar to those in the comparison schools in terms of prior attainment.

Additional findings

This evaluation found no evidence that the ScratchMaths intervention had an impact on pupil KS2 Maths attainment, measured at the end of Year 6. There was, however, a positive effect on computational thinking test scores at the end of Year 5. Exploratory analysis suggested that the size of this effect was larger for pupils who had ever been eligible for free school meals (everFSM), but did not differ between

boys and girls. This second finding is of interest as previous research has suggested that the impact of programming interventions can differ between genders.

There is evidence of poor implementation in intervention schools, particularly in Year 6. High fidelity to the programme was found in 44% of Year 5 and 24% of Year 6 classes. The process evaluation found that teacher attendance of training sessions, time spent teaching ScratchMaths and use of ScratchMaths materials all decreased between Year 5 and 6. 69% of survey responses from Year 6 teachers identified pressures around SATs as a barrier to implementation. Other reported barriers included the confidence and turnover of staff and the level of challenge of materials for the children. Exploratory analysis examined the impact of ScratchMaths only in those schools which delivered it with high or medium fidelity to the intervention but even for these schools no impact was found.

Teachers who sustained participation in terms of attendance and use of Scratch materials in accordance with trial protocols, were, in general, positive about the quality of the professional development and materials provided, particularly in Y5. Implementation was enhanced where schools provided time for teachers to work through materials. This was particularly so for teachers who were less familiar or unfamiliar with Scratch. It was also enhanced if schools showed a willingness to support computing teaching in Y6 despite the pressure of maths and English KS2 test requirements.

One of the issues explored as part of this evaluation was the relationship between computational thinking and mathematics attainment. The rationale of the programme is that improved computational thinking scores combined with teacher mediation (through the intervention) can lead to improvements in mathematics outcomes. However, although computational thinking scores did improve, it does not appear that the teacher mediation improved the translation of these skills into maths attainment: no evidence was found that the relationship between computational thinking and KS2 maths attainment was stronger within the schools that received ScratchMaths.

Cost

The average cost of ScratchMaths for one school was around £1,843, or £11 per pupil per year when averaged over 3 years. This does not include staff cover costs for the two days of professional development each participating teacher is required to attend each year. Future costs could be reduced by using local delivery partners for training instead of the single delivery partner in this trial (based at the UCL Institute of Education in London).

Table 1: Summary of impact on primary outcome

Outcome	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	P value	EEF cost rating
KS2 maths	0.00 (-0.12; +0.12)	0	🔒🔒🔒🔒🔒	5,818	0.970	£££££
KS2 maths (everFSM)	+0.01 (-0.14; +0.16)	0	N/A	1,632	0.915	£££££

Introduction

The intervention

Intervention description

ScratchMaths is a two-year computing and mathematics curriculum programme designed for pupils aged nine to 11 years, supported by teacher professional development. The ScratchMaths programme aims to address one difficulty many children have in learning mathematics - the need to express mathematical ideas in formal language. The development team's rationale for the ScratchMaths programme is in part to respond to that challenge, by finding a different language – and set of ideas and approaches - that are more open, more accessible and more learnable. At the same time, they sought to achieve this aim without sacrificing the rigour that makes mathematics work. Their hypothesis is that the language of programming can fulfil this role for pupils, providing that they work on carefully designed tasks and activities, and a teacher is able to support them.

In Year 1, the programme aims to enhance Scratch programming skills and computational thinking with connections made to areas of mathematics, and the materials and activities (detailed below) are geared towards that goal. Scratch is a freely available programming language, developed for educational purposes. It uses a visual interface and drag-and-drop tools. In the second year, as well as activities to develop computational thinking and Scratch programming, the content of materials also supports mathematical thinking more directly, through engagement with specially designed ScratchMaths curriculum activities and tasks linking programming to mathematical reasoning and understanding. Materials are aligned with both the Primary Computing (DfE 2013) and Primary Mathematics national curricula (DfE, 2014).

Teachers in the schools involved were offered 2.5 days of professional development per year.

The intervention was developed in 2014/15 during a 'design year' (see below). During the trial, Y5 teachers in participating schools- designated as 'Wave 1' - attended initial professional development in summer 2015 and the taught ScratchMaths to Y5 pupils in 2015/16. Y6 pupils attended professional development in 2016 and taught materials to Y6 pupils in 2016/17.

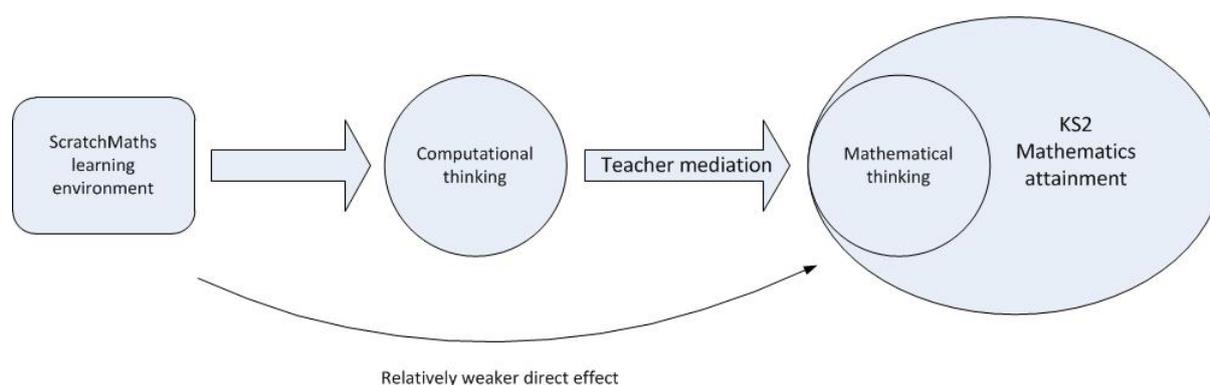
In addition, during 2016/17, in accordance with a waitlist design, Y5 pupils in control schools experienced ScratchMaths; these were designated Wave 2 schools.

The theory of change

The three figures below depict a theory of change in relation to the proposed intervention effect.

Figure 1 represents a simplified model of the hypothesised relationship between ScratchMaths, programming, mathematical thinking and KS2 mathematics attainment at the pupil level. Computational thinking, rather than programming and computing in general, is posited as the intermediate link between ScratchMaths and changes in mathematical thinking¹, thus an explicit focus on computational thinking informed the research questions and design. In any case, computational thinking and programming are generally considered as interlinked (see Brennan and Resnick, 2012; Selby, Dorling and Woollard, 2014; Meerbaum-Salant et al., 2013).

¹ See the evaluation protocol
https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_6-_Scratch_maths_amended.pdf

Figure 1: ScratchMaths and student learning

A revised version of the theory of change was proposed by the ScratchMaths team during the third year of the trial see Appendix C, figure 10². This revision suggested that the intermediate outcome is computing in general rather than computational thinking specifically.

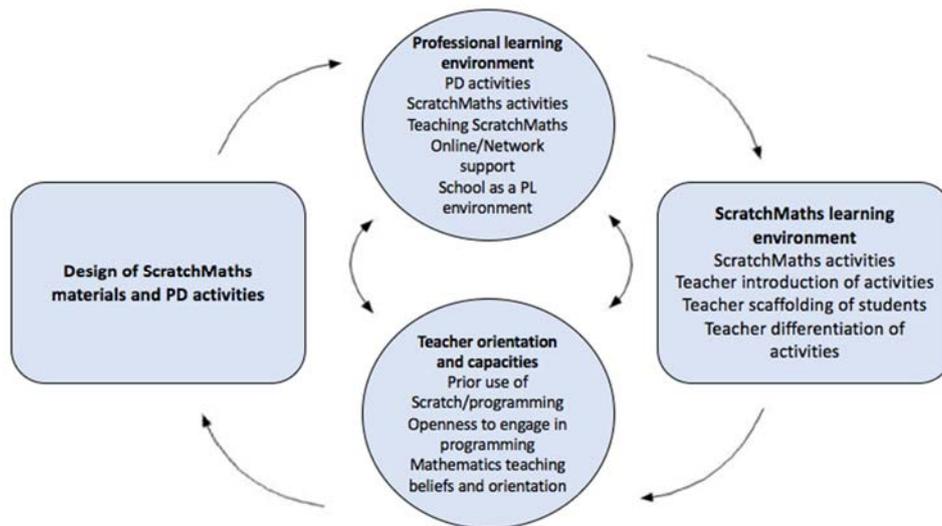
A critical feature of the theory of change is the importance of teacher mediation. Whilst engagement in Scratch programming may lead to improvements in computational thinking, the direct effect on mathematical thinking may be relatively weak. The ScratchMaths team also believe that, in general, enhancing mathematical thinking depends on teacher mediation and that ScratchMaths is no exception. For learning to happen, the teacher helps learners to make connections between computational thinking, Scratch programming and mathematics. Underlying the intervention design, teacher mediation is essential to foster learners' expression of mathematical thinking in the language of Scratch. The project design is focused on developing mathematical thinking, which is one aspect of the capacities and knowledge that are assessed through KS2 mathematics tests.

In addition to the posited theory of change in Figure 1, the ScratchMaths team also suggested a potential direct effect on KS2 mathematics (see evaluation protocol page 4). In discussion with the ScratchMaths team about the mathematical test used as a final measure, reference was made to the content of the 2017 KS2 maths test, the 2016 KS2 maths test and ScratchMaths content. ScratchMaths addresses or uses specific mathematics content and context, for example regarding angles. Thus, it might be expected that ScratchMaths would enhance learners' attainment in these specific content areas by providing opportunities for learners to explore specific concepts and to practise their existing knowledge in computing contexts. These effects would supplement any change in computational thinking or related mathematical thinking.

Figure 2 is a model of how the professional development and curriculum materials create the ScratchMaths learning environment and may lead to teacher change.

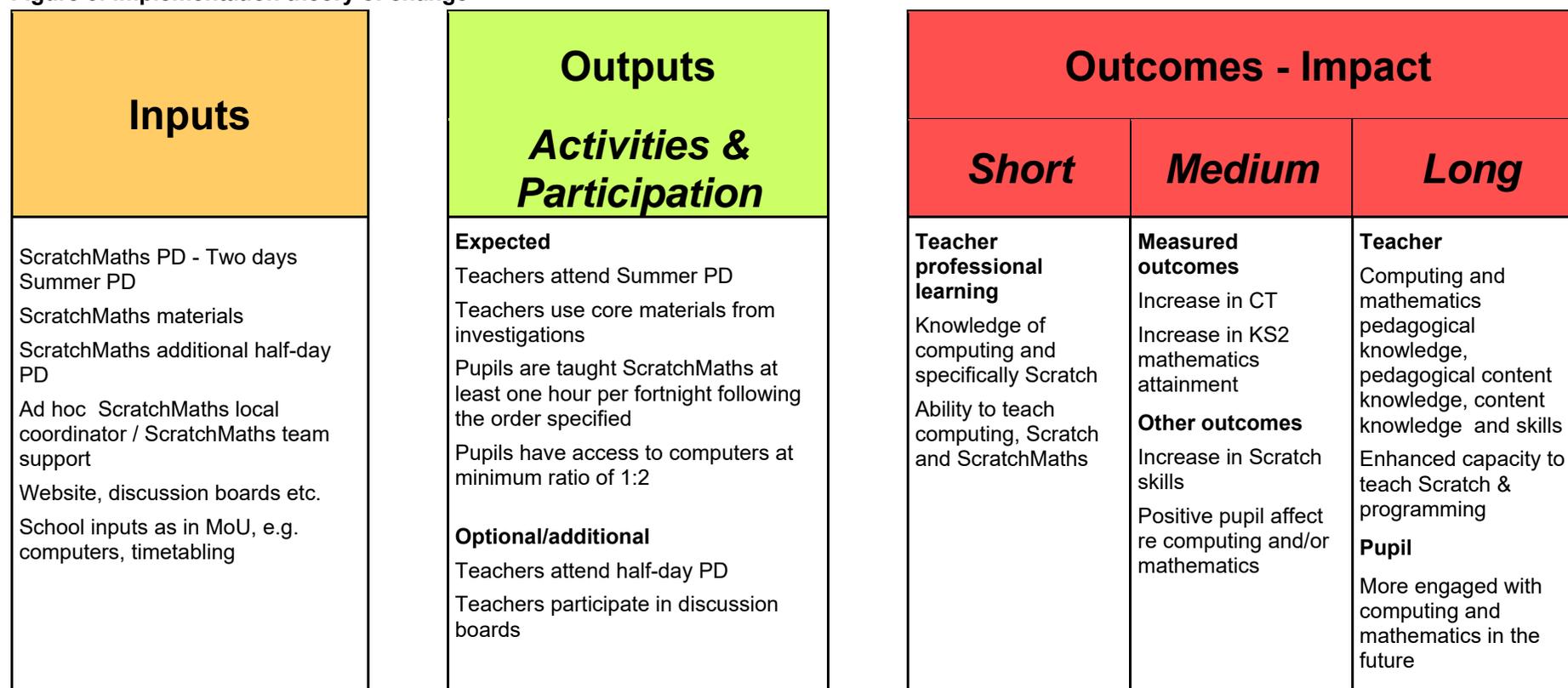
² In keeping with EEF guidance, the theory of change developed during the evaluation design is included in the report as this was the basis for the research questions and research design.

Figure 2: ScratchMaths as a professional learning system



The way in which professional development and teachers' engagement in curriculum innovation leads to changes in knowledge, practice and beliefs is more complex than this figure indicates. This is firstly because the professional learning environment and teachers' attitude to the intervention and their capacities are interconnected. Secondly, as can be seen from the diagram, the particular ScratchMaths learning environment a teacher instigates is itself an aspect of the teacher's professional learning environment. The model is of a set of *nested systems* (Opfer and Peddar, 2011). However, the limitations of the implementation and process evaluation of the trial mean that data have been collected on some aspects of the complexity of the professional learning system only. The implementation theory of change (logic model) is provided in Figure 3 below.

Figure 3: Implementation theory of change

**Assumptions**

The programme theory of change models the relationship between ScratchMaths, computational thinking and mathematical attainment.

Schools follow protocol/memorandum of understanding re: who teaches ScratchMaths and who engages with PD.

External Factors

Schools' capacity to participate (e.g. teacher release, suitable technology); school as PL environment; timetabling; who teaches; same/different teachers in Y5 and Y6; staffing changes; school leadership support/priorities; coherence/similarity of hubs; and hub leadership.

Prior experience of Scratch.

Recipients of the intervention

The recipients of the intervention were Y5 and Y6 pupils who experienced ScratchMaths materials in the context of the ScratchMaths learning environment. Teachers were the intermediary recipients of ScratchMaths professional development. The intended targets for recruitment were two-form entry schools, allowing for whole cohorts in each school to participate.

The target was for two teachers in each school to participate in professional development and teach ScratchMaths, and if possible these two teachers were Y5 teachers who would then be teaching the ScratchMaths curriculum the following year. Where they were unable to attend or it had not yet been confirmed who the Y5 teachers would be for the next school year, then one or more alternative teachers were asked to attend, such as the computing coordinator, who could share the training with other teachers in the school at a later date, or alternatively another class teacher (Y5 or Y6).

During the first year of the trial, in 2015-16, 2,986 Y5 pupils (9-10 years old) in 55 schools experienced ScratchMaths curriculum materials and activities. A total of 105 teachers in these schools attended at least some of the four professional development events. Although the intervention was designed for two teachers per school, in some schools only one teacher attended (if a one form entry) and in others more than two teachers attended, for example, with a specialist computing teacher and year group teachers attending, or if substitutes were sent in the case of illness.

Events offered consisted of two full days plus two half-days, with schools committing at recruitment for teachers to attend the full days, although this was not realised in all cases (see section on fidelity below). The majority of teachers attending the professional development events were Y5 teachers and then teachers who taught Y6 the following year (in most cases a different teacher). Attendees also included teachers with other roles, such as school computing coordinators.

In the second year of the trial, in 2016-17, the same classes of pupils progressing into Y6 (10-11 years old) experienced ScratchMaths. A total of 65 teachers, out of a target of 110 teachers, attended at least some of the 2.5 professional development days. A group of 24 teachers attended at least some of both the Y5 and Y6 training. In addition, in 2016-17, in accordance with a waitlist trial design, Y5 pupils in a further 55 designated control schools experienced the Y5 materials, with 64 teachers attending the professional development.

ScratchMaths materials

ScratchMaths materials are organised into three modules (per year). Modules develop knowledge and skills in relation to Scratch commands and concepts, computing concepts and mathematical content (see Appendix C for details). Each module consisted of a number of investigations, with four investigations developed for five of the modules and three for the sixth module (see Table 2 below).

An investigation consists of core activities that have certain steps designated as extensions, as well as some further, separate extension activities. In addition to programming activities to be performed on a computer, 'un-plugged' activities were included. Un-plugged activities are designed to develop computational thinking and programming skills through discussion, pen and paper activities and/or embodied activity away from the computer. Sets of investigations are brought together in a set of teacher materials for each module. Each investigation was designed to last in the range of 50-70 minutes (but in practice were likely to take at least two one-hour lessons, given the time needed for technical setup, and given that timing was also dependent on the number of extension steps/activities that a teacher chose to cover). One-off challenges were also available to be used either by individual pupils or with whole classes - these were additional activities and entirely separate from the core materials.

Building on Brennan and Resnick's (2012) emphasis on the importance of creativity as intrinsic to a computational perspective, the ScratchMaths team proposes a creative computing perspective in which

personal interests, agency and creativity are all important for meaningful engagement in the creation of computational artefacts. Within the ScratchMaths project, this is operationalised through a '5E approach': envisage, explore, exchange, explain, and extend (see Benton et al, 2017).

ScratchMaths professional development

Professional development in each year focused on the use of Scratch programming and the curriculum materials. Professional development took place in seven geographical 'hubs'.

Teachers from Wave 1 (intervention) schools attended two professional development days in the summer term prior to the teaching of materials to Y5 the following year. These days were intended to be obligatory. A gap between each professional development day was used to give participants opportunities between sessions to undertake tasks in school and to try out materials themselves. In the first year, there were two follow-up additional professional development half-days offered: one in autumn 2015 and one in spring 2016. The term 'additional' denotes that the half-day sessions were aimed at revisiting the professional learning covered in the full days, so attendance was not obligatory. Furthermore, in the second year, an additional half-day was offered in autumn 2016 and online webinars were offered in spring 2017.

During professional development events, participants engaged in the ScratchMaths activities as learners, were introduced to underlying concepts, and materials were introduced. The value of peer learning was modelled through paired discussion. Based on interview data, visits to events and survey data, the professional development was positively received on the whole.

Supporting materials and activities

Teachers' module materials included notes on using the materials, vocabulary and concepts, links to the primary national curriculum, and class discussion points. In addition, supplementary materials were provided, such as vocabulary sheets and Scratch starter projects. On the ScratchMaths website (<http://www.scratchmaths.org/wp-login.php>) teachers were able to access PowerPoint presentations that could be used to support the use of the investigations, short videos related to some activities, and could download starter code for Scratch projects. Table 2 provides details of the quantity of supplementary materials available for each module. In addition, an introductory video is provided on the ScratchMaths website for module 1.

Table 2: Module supplementary materials

Module	Scratch starter projects	Vocabulary sheets and/or posters	Presentations
1. Tiling Patterns	5	5	4
2. Beetle Geometry	6	5	4
3. Interacting sprites	4	4	4
4. Building with numbers	20	2	4
5. Exploring mathematical relationships	9	2	4
6. Coordinates and geometry	14	0	3

Intervention providers/implementers

For the intervention schools in both Year 1 and Year 2, the full-day professional development activities were led by the ScratchMaths team who had developed the curriculum materials, supported by

ScratchMaths local coordinators. At all professional development events led by the ScratchMaths team, at least two members of the team were present. Local coordinators had expertise in computing and/or mathematics teaching. They led one of the three optional half-day professional development sessions over the two years, with the ScratchMaths team leading on both the autumn term sessions. In addition, the ScratchMaths team led online webinars.

Location of the intervention

Schools were recruited in seven geographical clusters across England - Blackburn, Bradford, North London, South London, Merseyside, Somerset/Devon and Staffordshire. These seven areas were identified to represent a variety of different types of locality including urban areas and areas where schools were more dispersed and located in mixtures of urban and more rural locations. Identification of locations was done by the ScratchMaths team with advice from NAACE,

Adaptation

It was anticipated that the aims to recruit two-form entry schools and particular year group teachers for the professional development may need more flexible entry criteria. In relation to teacher use of materials, as described above, modules consisted of core materials, with extension activities and additional challenges. Thus, teacher selection of materials was envisaged. In addition, supporting materials such as teacher presentations were provided, but teachers were not obliged to use these and potentially could adapt them. Schools tended to run SCRATCH locally, rather than via the internet, due to technical issues.

Strategies to maximise effective implementation

During the first year of the project (2014/15), the ScratchMaths team from University College London, Institute of Education, undertook a development year in order to design and trial materials for the intervention (both Y5 and Y6 materials) and associated PD. It was planned that five 'design schools' would be involved. In the event, three were fully involved and two further schools had partial involvement. The design schools represented a range of school types and previous levels of engagement with Scratch programming. The pilot phase involved:

- Review of the literature and available materials.
- Collation, design and trial of materials in the design schools, culminating in a package of materials (teacher, student and PD) in preparation for the main trial focused on computational thinking (Y5) and mathematics and computational thinking (Y6).

The curriculum structure for both years of the intervention was developed prior to the commencement of the trial. Additionally, all of the Y5 content was designed prior to the Wave 1 (intervention) Y5 professional development, with a finalised version of all materials available online prior to the Wave 1 Y5 teachers commencing the delivery of the Y5 intervention in September 2015. Similarly, using the same initial structure, all of the Y6 content was designed prior to the Wave 1 Y6 professional development, again with a finalised version of all materials available online prior to the Wave 1 Y6 teachers commencing the delivery of the Y6 intervention in September 2016.

At the time of recruitment, all schools signed memoranda of understanding, agreeing to release teachers, and other requirements for participation (see Appendix D). ScratchMaths Local Coordinators (SMLCs) offered support to schools in their cluster by advising on recruitment processes and subsequently offering additional local support if requested.

Initially, it was planned to recruit schools to the trial in five geographical hubs, however this was changed to seven to support recruitment.

Implementation variability

Recruitment: the initial design aimed for recruitment of schools with two forms of entry for consistency of implementation. With two-form entry schools, and space for two teachers per year group available at PD events, a more consistent pattern of attendance would be possible. In addition, all pupils in the year group would experience ScratchMaths from teachers who had attended the PD events and so had direct and consistent PD experiences from the ScratchMaths team. However, the two-form entry criterion was relaxed to support recruitment (see 'Participant selection' in the Methods section below, for details of the recruited sample).

Professional development varied from the initial design, to include the use of a webinar for additional support rather than a spring twilight session. In addition, in the second year, a hub lead provided on-site professional development to a school. Other variability in implementation at school level and teacher level is reported below where process evaluation findings are reported.

Background evidence

Mathematics and computer programming in schools have a longstanding and intertwined history. There is evidence that programming in schools has the potential to develop higher levels of mathematical thinking in relation to aspects of number linked to multiplicative reasoning, and mathematical abstraction including algebraic thinking as well as problem-solving abilities (Clements, 2000).

More recently, attention has been paid to defining computational thinking (Brennan and Resnick, 2012; Cuny, Snyder and Wing, 2010; McMaster, Rague and Anderson, 2010; Selby and Woollard, 2013; Wing, 2008). Selby and Woollard (2013, based on a review of literature, provide the following definition of the characteristics of computational thinking:

- the ability to think in abstractions
- the ability to think in terms of decomposition
- the ability to think algorithmically
- the ability to think in terms of evaluations
- the ability to think in generalisations.

Computational thinking is posited to be a relative, or part, of the 'family' of different aspects of mathematical thinking (Wing, 2008). This relationship, if true, would help to explain why programming and programming-based mathematical learning have been found to have a positive effect both on student attitudes and attainment in mathematics in a meta-analysis of the effectiveness of computer-assisted instruction and computer programming in elementary and secondary mathematics (Lee, 1990). However, only a minority of studies in this meta-analysis focused on programming, and such studies were conducted in a period in which computers were more novel.

An agreed definition of computational thinking has not yet emerged in the literature, and the current programme of study in England for computing at KS2 makes no explicit reference to computational thinking (DfE, 2013). However, one of the aims of the computing programme of study addresses aspects of computational thinking:

[Pupils] Can understand and apply the fundamental principles and concepts of computer science, including abstraction, logic, algorithms and data representation (DfE, 2013)

Definitions are further discussed in Appendix F, where the development of the computational thinking test for the ScratchMaths trial is reported, along with descriptions of knowledge and skills developed in computing (Brennan and Resnick, 2012; Fuller et al., 2007; Meerbaum-Salant et al., 2013; ISTE and CSTA, n.d.).

The relationship between coding skills and computational thinking was explored in a recent randomised controlled trial (RCT) in England investigating Code Clubs (Straw, Bamford and Styles, 2017). Code Clubs are after-school clubs for 9 to 11 year-old children (the same age as those participating in the ScratchMaths trial). Teachers and volunteers support teaching of Scratch, HTML and Python programming languages. This RCT found positive outcomes in relation to children's programming skills but no significant effect on computational thinking as measured by a Bebras-based test (see Appendix F for discussion of Bebras).

A recent mixed-methods qualitative case study found a positive relationship between engagement in programming and problem-solving skills of 6 to 7 year-olds (Y2 in schools in England) when gender and prior attainment were accounted for (Blakemore, 2017). In the first phase of its introduction to the school curriculum, programming in schools was often developed by enthusiasts who would, in many cases, be located in mathematics departments, or, in primary school contexts, who would identify themselves as mathematics specialists. More generally, mathematics educators played an important role in promoting and developing programming, often linked to mathematical learning (see for example Hoyles and Noss, 1992). However, the introduction of Information and Communications Technology (ICT) as a national curriculum subject led to programming in schools being deprioritised, both in relation to computing and in the mathematics curriculum. Recent policy and curriculum changes mean that there is a renewed focus on computing being (re-)introduced into primary schools (DfE, 2013; Furber, 2012).

Since much of the research in computing and mathematical learning was undertaken, new programming languages (such as Scratch) and new tools (for example, Raspberry Pi) have been developed. Scratch is a freely available programming and authoring tool developed for educational purposes at the Massachusetts Institute of Technology in the early 21st century (Monroy-Hernandez and Resnick, 2008). Scratch is based on graphical programming blocks that can be assembled to create programs. The appearance of the blocks is related to their function and meaning. Scratch is highly interactive, allowing for changes to be made to programs while they are running. The scripting area is modelled on a physical desktop to encourage 'tinkering'. Scratch is designed to interface with multimedia, allowing for applications that are meaningful to learners (Resnick et al., 2009). Computational knowledge and skills developed in Scratch have been identified (Brennan and Resnick, 2012).

While Scratch is a development of earlier programming languages designed as learning environments, it represents a significant change in how users code and develop conceptual understanding of programming. Thus, the ScratchMaths intervention represented an opportunity to design and evaluate a curriculum and professional development programme aimed at maximising the benefits of programming for students' mathematical thinking and attainment in the current context.

Evaluation objectives and research questions

Objectives

Evaluation objectives were specified prior to the trial in the evaluation protocol³. The impact evaluation sought to:

- Identify the effect of the intervention on mathematics attainment.
- Through the design of a computational thinking assessment, to establish the effect of the intervention on computational thinking as a secondary, intermediate measure, as well as the relationship between computational thinking and mathematics attainment.
- Provide an independent view on the process of the design of the curriculum materials and associated professional development activities and the ScratchMaths team's evaluation of

³ https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_6-Scratch_maths_amended.pdf

these, and to provide guidance to the team on ensuring that the intervention approaches, materials and training will be replicable and testable through a randomised control trial.

- Evaluate the reliability and validity of any identified impact through a process evaluation designed to identify issues of fidelity and scalability, in particular the barriers and necessary conditions for successful implementation, and to address the evaluation research questions.

Research questions

RQ1: What has been the effect of the intervention on the development of pupils' mathematical skills as measured by a randomised control trial?

RQ2: How can computational thinking be measured?

RQ3: What correlation exists between measured computational thinking and mathematics attainment?

RQ4: What has been the impact of the intervention on the development of pupils' computational thinking?

RQ5: What conclusions can be drawn about the relationship between mathematical thinking and computational thinking from the quantitative analysis?

RQ6: To what extent does the design and delivery of curriculum materials and professional development and the associated materials fit with the current knowledge base on effective professional development in relation to mathematics teaching/computing?

RQ7: What are the teachers' views on the effectiveness of the professional development?

RQ8: Were there any barriers to implementing Scratch, or were there particular conditions that needed to be in place for it to succeed?

RQ9: In what ways can the professional development delivery and materials be improved?

RQ10: What issues of fidelity occur during the trial, and how secure is the trial outcome (taking into account any use of Scratch in control schools)?

In addition to these research questions, there was also an exploration, through the process evaluation, of the scalability of the trial.

Ethical review

The trial received ethical approval through both Sheffield Hallam University's and Institute of Education's (University College London's) ethics processes. National Pupil Database (NPD) data were subject to NPD protocols about data sharing. At both institutions, procedures are in place to comply with the 1998 Data Protection Act and, to the best of our knowledge; both universities conform to the principles of ISO/IEC 27001 information security standards. No information about any identified individual was reported or made available to anyone beyond the project teams. All data were stored anonymously and securely. Consent forms, participant information, and digital recordings have all been stored separately from any transcripts and case reports. In disseminating findings, names of respondents appear as pseudonyms, and any other potentially identifying data are anonymised. Personal data are only stored on encrypted portable media in password-protected files (and only when absolutely necessary).

Prior to randomisation, all schools applying to take part in the trial provided a memorandum of understanding (MoU) to the ScratchMaths IoE team, with a copy later provided to Sheffield Hallam University (SHU). The MoU was signed by the headteacher (see Appendix D). Opt-out parental consent

was obtained with a total of 21 pupils opting out⁴ (see Appendix D). For telephone interviews with participating teachers, additional information was provided and written consent obtained. Teacher survey participants were informed that completion and submission of surveys constituted consent.

Project team

Firstly, the ScratchMaths development team and, secondly, the SHU evaluation team members are listed below. For a fuller description of roles see Appendix E.

ScratchMaths development team, University College London, Institute of Education (IoE)

Professor Richard Noss, Professor Celia Hoyles, Professor Ivan Kalas, Professor Dave Pratt, Laura Benton, Alison Clark-Wilson, Kim Parsons, Piers Saunders, Johanna Carvajal.

National Association for the Advancement of Computer Education (Naace)

Mark Chambers: CEO.

Evaluation team, Sheffield Hallam University (SHU)

Professor Mark Boylan, Sean Demack, Dr John Reidy, Anna Stevens, Claire Wolstenholme, Dr Martin Culliney, Ian Guest, Professor Hilary Povey, Phil Spencer, Sarah Reaney-Wood, Ian Chesters.

Trial registration

The Scratch Maths trial was registered on 11th October 2016 on The ISRCTN registry (ID number ISRCTN10189078)⁵.

⁴ Fifteen pupils opted out prior to randomisation and six pupils opted out during the first year of the evaluation. No pupils opted out during the second year of the evaluation.

⁵ See <http://www.isrctn.com/ISRCTN10189078>

Methods

Trial design

The impact of ScratchMaths was evaluated using a two-armed clustered randomised controlled trial (RCT) design with randomisation at the school level. Randomisation took place at the school level in order to minimise the risk of spill-over that within-school randomisation brings. The RCT design took account of clustering of pupils within schools in classes and clustering of schools within seven geographical areas⁶.

In April 2015, 110 of the 111 recruited schools were randomised into the intervention (Wave 1) or control (Wave 2) groups. For the control condition, a waitlist approach was adopted, as detailed below:

Intervention (Wave 1) Schools (2,986 pupils in 97 classes in 55 schools):

- ScratchMaths professional development events for Y5 teachers in summer 2015 and Y6 teachers in summer 2016.
- Schools implementing ScratchMaths with Y5 classes in 2015/16 and with Y6 in 2016/17.

Control (Wave 2) Schools (3,246 pupils in 110 classes in 55 schools):

- ScratchMaths Professional Development events for Y5 teachers in summer 2016 and Y6 in summer 2017 (after trial end).
- Pupils/teachers in Y5 during 2015/16 and Y6 in 2016/17 represent the 'business as usual' control group.

It is important to note the distinction between class and teacher level. Prior to randomisation in April 2015, recruited schools submitted class lists for all Y4 pupils. The numbers of pupils, classes and schools at baseline (shown above for intervention and control schools) are based on this data. The ScratchMaths program was aimed at two teachers in each intervention school - two teachers in Y5 during 2016 and Y6 in 2017. The data provided prior to randomisation was insufficient to attach named teacher(s) to specific Y4 classes. So, whilst this trial had a class level, this does not mean it had a teacher level.

The ITT impact analysis sample is based on these submitted class-level lists of all Y4 pupils (who had complete baseline/outcome data and did not opt out) for each school. The number of classes per school varied between 1 and 4 (see table 3, below). The 55 intervention schools had a mean of 1.75 classes per school and a mean of 30.8 pupils per class. The 55 control schools had a mean of 2.00 classes per school and a mean of 29.5 pupils per class. This accounts for the slightly larger pupil sample in control schools (n=3,246) compared with intervention schools (n=2,986).

As detailed in the evaluation protocol⁷ and randomisation section below, a propensity-score-paired-school-stratification approach to randomisation was adopted. This approach grouped schools into 'nearest statistical' pairs within the geographical hub regions. For each pair, one school was randomly selected into the intervention group and the other was placed into the control group.

KS1 National Pupil Database (NPD) data for attainment, Free School Meal (FSM) status and gender were obtained in January 2016. These KS1 data were collected in 2013, two years prior to randomisation. NPD data for the primary outcome (KS2 maths attainment) were requested in July 2017

⁶ As noted above - Blackburn, Bradford, North London, South London, Merseyside, Somerset/Devon and Staffordshire.

⁷https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_6-_Scratch_maths_amended.pdf

and obtained in December 2017. Data for the interim secondary outcome (computational thinking) were collected in July 2016.

Note that the waitlist design in this case meant that Y5 pupils in the control schools received the intervention during the second year of the trial (2016/17). There is, then, a risk of potential spill-over from those Y5 teachers and classes to the control Y6 teachers and classes. Data investigating the possible spill-over were collected as part of the implementation and process evaluation (IPE) by a survey of teachers.

Two key documents were published on the EEF website during the trial period: the evaluation protocol and the statistical analysis plan (SAP). The evaluation protocol was first published in February 2016⁶ and updated in March 2017 and the SAP was published in November 2017⁸.

Participant selection

Eligibility criteria

The trial was to involve, ideally, two classes of Y5 pupils per school, and the same pupils when they progress into Y6. The target recruitment was for two-form entry schools wherever possible. Having the same number of Y5 classes would simplify Unique Pupil Number (UPN) data collection, opt-out consent and possible movement between classes as the whole year group would be involved. The aim for two-form entry schools was relaxed, due to recruitment difficulties, in order to support recruitment, and the distribution of classes in schools is set out below in Table 3.

Table 3: Sample by number of forms (classes) per school

Number of forms	Number of schools: Intervention	Number of schools: Control	Total
1	5	2	7
1.5	1	0	1
2	43	43	86
3	6	8	14
4	0	2	2
Total	55	55	110

Data source: table compiled from number of forms indicated by information supplied by school to IoE, with missing data completed from Raise Online⁹ as source for 10 of the 110 schools.

There is the possibility of a small dilution effect for the six three-form entry intervention schools (though arguably a balancing concentration effect in one-form entry schools). School size and number of forms of entry are included in the intention-to-treat analysis model through the matching process.

Other criteria were that schools had adequate internet connectivity and enough laptops or desktop PCs (at least one between two pupils) available for Y5 and Y6 pupils.

⁸ See <https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/scratch-programming/> for these documents.

⁹ <https://www.raiseonline.org/>

Recruitment

Recruitment began during the design phase with the aim of identifying all schools by 15 March 2015. This was to allow for signing of MoUs, collection of UPNs and NPD data retrieval prior to randomisation. IoE aimed to identify five hub locations through discussion with the National Association of Advisors for Computers in Education (UK) (Naace). In the event, seven hubs were identified that were geographically spread and had different profiles in terms of degrees of urban or rural contexts. Recruitment was undertaken by IoE with support from Naace. Once schools had been recruited, SHU collected UPNs for the focus cohort (Y5 in 2015/16) and other school and teacher-level data as needed.

The trial planned to involve the recruitment of approximately 115 schools. It was anticipated there would be drop-out at the point of agreeing to trial protocols, and the design aimed for 100 participating schools at the point of randomisation (50 intervention, 50 control). In the event, 111 were recruited and 110 allocated to the intervention and control conditions (see below). As noted above, recruitment was undertaken by the ScratchMaths team working with the Naace and local partners and so contact with schools was undertaken through multiple pathways and therefore it is not possible to identify how many schools were approached to take part.

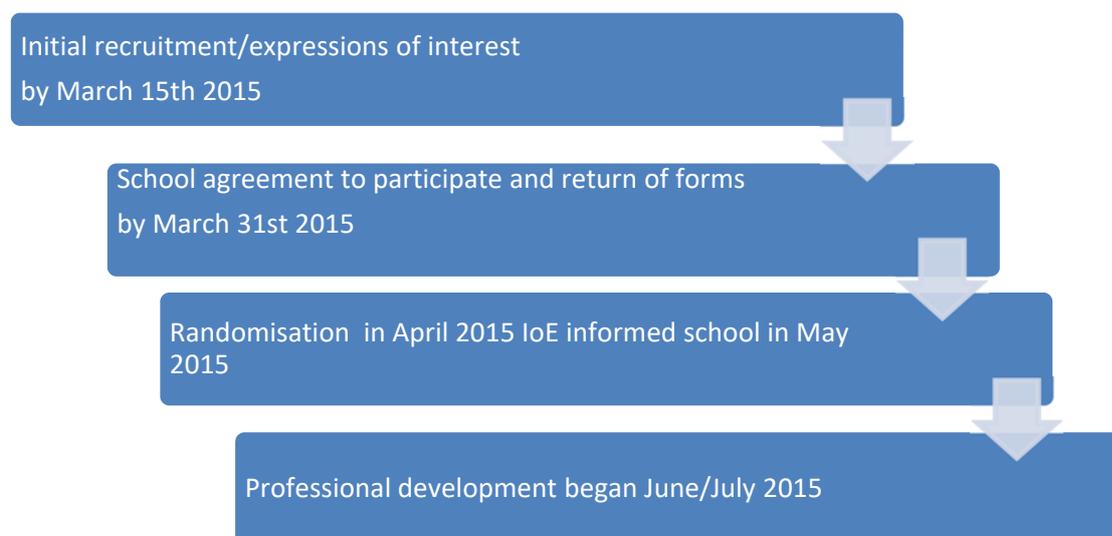
SHU and IoE co-produced information, including initial recruitment information, consent forms and MoUs for use with the schools. IoE supplied SHU with information on the recruitment process using the participant flow diagram recommended for EEF reports¹⁰.

Schools were required to provide the following as a condition of being entered into the randomisation:

- MoU signed by the head teacher. The MoU included details of the requirements for the computational thinking test (CT test) in summer 2016 and both IoE and SHU evaluation activities, as well as information the school would be expected to supply.
- Information on names of teachers and roles of those who would be attending the professional development events if allocated to the intervention group.
- Summary information on any previous use of Scratch programming, or engagement with Bebras/Beaver tests (this information was collected for purposes of comparing samples following randomisation but was not related to eligibility criteria).
- Pupil lists for Y5, including UPNs.
- Confirmation the school has sent out the parent opt-out consent form.

Figure 4 outlines the process of recruitment, with dates, up until the start of professional development.

¹⁰ <https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluating-projects/evaluator-resources/writing-a-research-report/>

Figure 4: Recruitment and allocation timeline**Memorandum of understanding and consent**

In addition to being sent a project information sheet with ethics and consent issues outlined, all schools were asked to sign and return (by the head teacher) an MoU (see Appendix D) which detailed the school's, the ScratchMaths development team's and evaluation research team's responsibilities to the project. The MoU acted as a contract between the school and the research team, and enabled participants to understand and give their consent to all aspects of the trial. The MoU explained that the research team would access the NPD to retrieve data on pupils involved in the trial. Data collected from the test and schools would then be matched with data retrieved from the NPD and could be shared with IoE, the EEF's data archive and the UK Data Archive for research purposes. Lastly the MoU explained that no individual school or pupil would be identified in any report arising from the research. Similar procedures were used in schools participating in the CT test development (see Appendix F).

Consent for pupils to take the CT test (in both design and trial schools) was obtained through opt-out parental consent forms distributed to all parents of pupils taking the test via the school (see Appendix D for main trial form). A total of 15 parents completed an opt-out consent form in the first instance (before the NPD request was submitted). After class lists were obtained, a further six parents returned opt-out forms. The total number of pupils for whom baseline data was considered, excluding opt-outs, is 6226 (6232-6).

All teachers in the intervention and control samples, who were asked to complete a survey, when contacted by email, were sent an information sheet which detailed ethical procedures including data storage and usage. Teachers gave their consent for use of their data through their completion of the survey and this was outlined in the information on the first page of the survey. Teachers taking part in a telephone interview were also sent the project information sheet prior to arranging an interview. Ethical procedures were discussed at the start of the interview, including their right to withdraw, and consent was taken verbally from teachers to take part in the interview and for it to be recorded.

Outcome measures**Primary outcome**

The primary outcome was overall KS2 maths attainment in May 2017. Appendix G summarises the distribution of two measures of KS2 maths attainment: a raw score (obtained from summing the scores in three KS2 maths papers) and a score that is re-scaled such that a value of 100 or greater indicates

when a pupil has met or exceeded the expected KS2 maths level of attainment¹¹. The SAP specified the primary outcome for the impact analysis to be the raw KS2 maths attainment score. However, because the distribution of the raw scores displayed a notable skew, all of the intention-to-treat (ITT) impact analyses were replicated using the scaled (and not skewed) version of KS2 maths attainment. See Appendix G for more detail on this.

Secondary outcomes

Follow-on analyses examined impact within the three KS2 maths test papers¹² taken by pupils in May 2017.

In addition to KS2 maths attainment, a further secondary outcome was computational thinking based upon pupil scores for a CT test developed administered to trial participants in July 2016. Thus, this measure also provided an interim measure of the impact of ScratchMaths at the end of the first year of the trial. The CT test was piloted by SHU in 2015, and this pilot and the use of the test in the main trial addressed Research Question 2 "how can computational thinking be measured?" Further details of the CT test, its development and analysis are provided in Appendix F and are summarised below.

Members of the SHU and IoE teams met in November 2014 for SHU to develop an understanding of IoE's operational definition of computational thinking as used in the intervention design. Following this, SHU reviewed the literature on computational thinking, and proceeded to design, develop and test the CT test independently and prior to having access to ScratchMaths materials. Further, the ScratchMaths team had access to the test only once it had been used in the main trial and after material development. The CT test used Beaver/Bebras¹³ questions or similar types of tasks, to support construct validity. The selection of items was informed by the composition and level of difficulty of English versions of Beaver/Bebras designed for 10-11 year-olds. The test and test protocols were piloted with both Y5 and Y6 children across the attainment range and then administered to an outcome measure design sample of 231 Y6 pupils from nine primary schools in England from a region not involved in the trial. The CT test scores from these pupils were then correlated with KS2 maths scores and this yielded a statistically significant correlation ($r=0.45$, $n=231$, $p<.001$).

Key features of the CT test, as identified from both the pilot and main trial samples were:

- High level of construct validity given the use of Beaver/Bebras and similar items.
- Normal distribution of scores with a mean near to the mid value of the scale.
- Good internal reliability (ordinal Cronbach's α of 0.72).
- Unidimensional.¹⁴
- All items had significant factor loadings with the single underlying dimension and all items if deleted led to a reduction in ordinal Cronbach's α .

The CT test was developed independently of the ScratchMaths team. Following its use in the main trial, the ScratchMaths team expressed concerns about the validity of the CT test in relation to it focusing on 'pre-formal' aspects of computational thinking, and it not being related to the new computing national curriculum. These concerns are included and discussed in Appendix F alongside limitations identified by the evaluation team.

ScratchMaths may have a positive impact on aspects of computational thinking that are not tested by the CT test, and it is important to recognise that it is not a test of programming knowledge or skill.

¹¹ See <https://www.gov.uk/guidance/scaled-scores-at-key-stage-2>

¹² Specific NPD variables - KS2_MATARITHMRK (Paper 1, arithmetic); KS2_MATPAPER2MRK (Paper 2, reasoning) & KS2_MATPAPER3MRK (Paper 3, reasoning).

¹³ Established in 2004, this is an international computing contest, see <http://www.bebbras.org> and <http://www.beaver-comp.org.uk>. It is now run in 30 countries. In 2013, more than 720,000 pupils took part in the contest.

¹⁴ Factor analysis using parallel analysis indicated one dimension and Rasch analysis yielded a non-significant Andersen Likelihood Ratio test ($\chi^2 = 4.14$, $df = 9$, $p = .902$) also suggesting unidimensionality.

However, there are sufficient reasons to conclude that it was suitable for use in the trial, given that: the overall outcome of normal distribution around the middle of the test scale (see Appendix H); the test did identify a difference in outcome; and it has a high level of construct validity as the type of items, mostly from Beaver/Bebras, are ones considered to be related to computational thinking and of the level of difficulty considered suitable for children of this age.

Measuring computational thinking at the end of Year 1 (2015/16) addressed the intended outcomes of the first year of the trial, where the focus was on computing and computational thinking. In addition, it allowed analysis of the relationship between measured differences in computational thinking and the impact on mathematics attainment.

Administration of the CT test in the main trial was staggered to accommodate schools' access to IT facilities and the potential need for support with log-in for pupils and IT support by SHU. Appendix F has more details about the timing of the CT test. As approximately 75% of school tests took place within a two-week period, it is unlikely that timing influenced outcomes, and there were no discernible differences in patterns of test-taking between intervention and control samples.

Assessment was automated through application of code to student responses and so was blind to the trial condition. Teachers were responsible for invigilating the tests, following a protocol developed during the pilot of the test; they were also required to submit a record of any relevant factors that might have influenced results. Teacher invigilation of the test had the same level of security as routine administration of KS2 tests. Analysis of records of test administration from teachers indicates no threats to test reliability from the way the test was administered, comparing the intervention and control conditions.

Knowledge of outcomes of the comparative analysis of the CT test was restricted to the trial statisticians and so was withheld from schools, IoE and other members of the SHU evaluation team, until summer 2017, when the trial was complete.

Sample size

A three-level clustered randomised controlled trial design was adopted for this evaluation (pupils clustered into classes clustered into schools). Randomisation took place at the school level and the outcome variables are all at the individual pupil level. A class-level analysis was also included to reflect the structural reality of the data and to acknowledge the widespread use of setting within primary schools for KS2 mathematics.

The power calculations were undertaken using the Optimal Design Software¹⁵. Table 4 below summarises the estimated minimum detectable effect sizes (MDES) for the primary and secondary outcomes from the protocol, baseline and at analysis stages of the trial.

At the protocol stage, an MDES of 0.18 standard deviations was estimated. Specifically, a three-level clustered randomised controlled trial with 110 schools, two classes per school and 20 pupils per class, results in a design that would be able to detect an effect size of 0.18 standard deviations or higher as statistically significant with a statistical power of 0.80. This estimate assumed that an estimated 13% of the variation in KS2 maths attainment would be clustered at the school level and 7% would be at the maths classroom level¹⁶. Additionally, the protocol MDES estimate assumed that the correlation between KS1 and KS2 maths attainment would be 0.77 ($R^2=0.59$) based on EEF guidance. The baseline numbers for schools and classes were very similar to those we had estimated in the protocol,

¹⁵ Raudenbush, S. W., et al. (2011). *Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01)* [Software]. Available from www.wtgrantfoundation.org

¹⁶ The school level ICC of 0.13 is taken from the EEF guidance from analyses of NPD 2013-2014 and the class level ICC of 0.07 is estimated as being half of what is found at the school level (due to the widespread practice of setting within primary mathematics).

while the number of pupils was larger. This resulted in the same MDES estimate at baseline as was reported in the protocol.

For the interim secondary outcome (computational thinking), the protocol MDES estimate was 0.21 standard deviations, which also remained consistent at baseline.

Table 4: Minimum detectable effect size (MDES) for planned analyses for ScratchMaths clustered RCT from protocol and at baseline

	Protocol	Baseline
No. of Schools / Classes / Pupils	110 / 220 / 4,400	110 / 207 / 6,232
Primary Outcome (KS2 Maths Attainment)		
KS1 Maths Covariate R ²	0.59	0.59
MDES estimate	0.18	0.18
Interim Secondary Outcome (CT Test Score)		
KS1 Maths Covariate R ²	0.25	0.25
MDES estimate	0.21	0.21

p<0.05; statistical power =0.80; ICC estimates: 0.13 (school); 0.07 (class)

Randomisation

A stratified approach was adopted for the school-level randomisation in April 2015. As detailed in the trial protocol and SAP, a logistic regression model was used to generate school-level predicted probability or 'propensity' scores based on the schools' 2013/14 KS2 attainment outcome variable¹⁷ and seven explanatory variables¹⁸. Within each of the hub areas, the propensity scores were used to group schools into their 'nearest statistical neighbour' pairs. One school from each pair was then randomly selected into the intervention (Wave 1) group; the remaining school was allocated to the control (Wave 2) group.

Our propensity-score-paired-school-stratification approach required an even number of schools in all of the geographical hub areas. This was not the case for three areas: an odd number of schools were recruited in the two London hubs (north and south) and in the Somerset hub. The two London hubs were merged into a single hub with an even number of schools. Within the Somerset area, the propensity scores identified one school to be very distinct¹⁹ from the remaining 16 schools. This school was then dropped and the remaining schools were paired and randomised to the intervention or control group. The Somerset school excluded from the trial was offered ScratchMaths as part of the waitlist design, but we have not used any data from this school in the impact evaluation.

In all, 55 schools were randomly selected to receive the ScratchMaths intervention and their 55 paired schools were allocated into the control group.

¹⁷ A binary outcome that identified whether the proportion of pupils within a school attaining a level 5 or higher in KS2 mathematics was greater than the median population value of 42% (=1) or not (=0). Source: KS2 School level Census data for all of England in the 2013/14 academic year, available from <https://www.compare-school-performance.service.gov.uk/download-data>

¹⁸ Explanatory variables - KS1 attainment, KS1 to KS2 progress in mathematics, school size, gender balance, %FSM, %EAL, %SEN.

¹⁹ In terms of propensity scores, this school had a score of 0.403. When the propensity scores were rank ordered within the Somerset area, the score for the school immediately below was 0.303 and the score for the school immediately above was 0.949. The school immediately below was paired with a school with a score of 0.297 and the school immediately above was paired with a school with a score of 0.960.

In 2014/15, prior to randomisation, recruited schools were requested to provide lists of all pupils in Y4 and the names of their classes and teachers. Following randomisation in April 2015, 2,986 pupils were located in 97 classes in the 55 intervention schools and 3,246 pupils located in 110 classes in the 55 control schools.

The propensity-score-paired-school-stratification approach to randomisation brings three key advantages:

- It enabled a large number of variables to be drawn on for stratification.
- Stratification variables were finely grained (i.e. scale rather than categorical).
- It provided greater flexibility and robustness for follow-on analyses (such as on-treatment or sensitivity analyses).

The propensity-score-paired-school-stratification approach drew on seven scale variables to create 55 pairs of schools. Within each pair, one school was randomly selected into the intervention (Wave 1) and the other to the control (Wave 2) group. Minimisation and standard approaches to stratification tend to be confined to a smaller number of categorical variables and therefore are less finely grained.

The third advantage is most clearly illustrated with respect to on-treatment analyses. In the ScratchMaths trial, an on-treatment analysis might proceed from analyses that examined whether fidelity to ScratchMaths was statistically associated with KS2 maths attainment. If an association between fidelity and attainment was found, a subsample of the intervention group might be identified as being 'on treatment' to ScratchMaths if they are observed to reach a specified level of 'fidelity'.

An on-treatment impact analysis might then compare the KS2 maths attainment for this restricted 'on treatment' intervention group subsample with a control group. With minimisation or more standard stratification approaches to randomisation, the original complete control group would usually be used for this comparison. This brings an increased risk of imbalance between the restricted 'on-treatment' intervention subsample and the original complete control sample.

The propensity-score-paired-school-stratification approach limited this risk of imbalance. Once an 'on-treatment' intervention subsample of schools is identified, the control group can similarly be restricted to include just the matched pairs for each 'on-treatment' school. For example, if 30 of the 55 intervention schools involved in the trial are identified as 'on treatment', these 30 intervention schools could be compared with their 30 matched control schools rather than the entire 55 control school sample. This is the approach that was followed in the analysis.

Analysis

Impact analyses for primary outcome

As set out in the SAP, the impact of ScratchMaths on KS2 mathematics attainment was examined using a multilevel analysis with three levels (pupils clustered into classes²⁰ clustered into schools). An intention-to-treat (ITT) approach was adopted and the models were constructed in three stages.

- Stage 1 - an outcome-only analysis that included the dummy variable that identified whether a pupil was in the intervention or control group.
- Stage 2 - KS1 maths covariates at both pupil and school level²¹ were included.

²⁰ See Trial Design in Methods section above, note that class and teacher levels are not the same. The class level was identified using class list data provided by recruited schools prior to randomisation.

²¹ KS1 mathematics attainment was included as a pupil level (NPD variable name = KS1_MATPOINTS) and an aggregated (mean pupil score) version was included at the school level.

- Stage 3 - all of the school-level variables used as explanatory variables²² to generate the propensity scores and dummy variables to identify school pairs within geographical hubs were included.

The Stage 2 model was used to assess the impact of the ScratchMaths intervention on the primary outcome. The impact of ScratchMaths has been estimated by dividing the coefficient for the dummy variable that identifies ScratchMaths intervention schools by the total standard deviation for the empty multilevel model (see Appendix I for more detail on this). The Stage 3 model was undertaken as sensitivity analysis to fully take account of the propensity-score-paired-school-stratification research design.

To address a notable negative skew observed in the primary outcome²³, further sensitivity analyses were undertaken not specified in the SAP. A scaled version of KS2 maths attainment was supplemented for the specified raw KS2 maths outcome and modelled using the same three stages listed above.

As specified in the SAP, three subsample analyses were undertaken. The purpose of these analyses was to explore whether ScratchMaths had a different impact for some groups of pupils compared with others. Subsample analyses relating to FSM status, gender and KS1 maths attainment were undertaken. This was done first by introducing interaction terms into the KS2 maths attainment impact model using two stages:

- Stage 1 - Main Effects model that included the ScratchMaths dummy variable, KS1 maths attainment, FSM status and gender.
- Stage 2 - including interaction terms ScratchMaths*FSM; ScratchMaths*Female and ScratchMaths*KS1 maths alone and then simultaneously.

Follow-on subsample analyses for FSM and not-FSM subsamples were undertaken regardless of the findings from the interaction analyses (FSM being a sub-group of interest for the EEF), but follow-on subsample analyses relating to gender and KS1 maths attainment were only undertaken if the interaction term was statistically significant.

Impact analyses for follow-on KS2 maths secondary outcomes

As specified in the SAP, follow-on analyses examined the impact of ScratchMaths on attainment within the three separate KS2 maths tests. These analyses adopted exactly the same ITT approach used for the main primary analyses and exactly the same three-level and three-stage multilevel approach.

Impact analyses for interim computational thinking test outcome (2016)

An ITT approach for the interim secondary outcome (computational thinking) was not possible due to issues of missing data. CT test data were not obtained from 29 schools including 11 intervention schools known to have withdrawn from engagement in ScratchMaths professional development or use of materials²⁴. Reasons for non-completion of the other 18 schools are not known. However, a similar number of control schools did not participate in the wait-list PD events and so this may signify that they felt less investment in participating. The main impact analyses for the CT test outcome adopted the same three-level and three-stage multilevel approach taken with the primary outcome.

In response to the potential imbalance that missing CT test data from 29 schools might bring, further sensitivity analyses were conducted. These analyses drew on the propensity-score-paired-school-

²² School-level variables - KS1 attainment, KS1 to KS2 progress in mathematics, school size, gender balance (% Female), %FSM, %EAL, %SEN.

²³ The negative skew was observed in the KS2 maths raw attainment score (KS2_MATMRK), see Appendix I.

²⁴ The total of 11 is based on information from the ScratchMaths team or provided by schools when asked to undertake the CT test. The situation of a further four schools is ambiguous.

stratification research design to limit the CT test analyses to a sample of 'complete pair' schools. Specifically, within the 81 schools with CT test data, 62 were 'complete pairs' and CT data were available for both the intervention and paired control schools (31 intervention and 31 matched control schools). As discussed in the SAP and summarised in the trial design section above, the reason for doing this was to best ensure good baseline balance between control and intervention schools without compromising randomness in the RCT design.

As with the primary outcome, subsample analyses relating to FSM status, gender and KS1 maths attainment were undertaken for the CT test outcome. This was done using two model stages:

- Stage 1 - main effects model that included the ScratchMaths dummy variable, KS1 maths attainment, FSM status and gender.
- Stage 2 - including interaction terms ScratchMaths*FSM, ScratchMaths*Female and ScratchMaths*KS1 Maths alone and then simultaneously.

Follow-on subsample analyses to explore the impact of ScratchMaths on the CT test outcome for FSM and not-FSM subsamples were undertaken regardless of the findings from the interaction analyses. Follow-on subsample analyses relating to gender and KS1 maths attainment were only undertaken if the interaction term was observed to be statistically significant.

On-treatment analyses for primary outcome

In discussion with IoE, fidelity to the ScratchMaths intervention was defined in terms of five dimensions set out in Table 4 of the SAP. Specifically, the five (school-level) dimensions were: attendance of ScratchMaths PD events; pupils' access to computers; coverage of ScratchMaths modules; ScratchMaths curriculum time; and order/progression of ScratchMaths modules. Appendix K draws the five ScratchMaths fidelity dimensions together to identify a sample of pupils located in five schools that were identified as having high fidelity to ScratchMaths over the two-year trial period. In Appendix K, a sample of pupils located in 13 schools identified as having medium or high fidelity is also identified.

On-treatment analyses for the primary outcome were undertaken using model stages 1 and 2 of the main ITT impact analyses. In the on-treatment analyses, the intervention group sample was restricted to pupils in the five high-fidelity or 13 medium/high-fidelity intervention schools. The KS2 maths attainment for pupils in the restricted fidelity intervention samples were compared with the attainment for pupils in control schools. This was done first using the raw control sample of 55 schools and then drawing on the propensity-score-paired-school-stratification research design to limit the control samples just to those that were matched to the five high-fidelity or 13 medium/high-fidelity intervention schools prior to randomisation.

Follow-on exploratory analyses for primary outcome

Further follow-on analyses explored the relationship between ScratchMaths, computational thinking and KS2 mathematics attainment. This was done using the following three model stages:

- Stage 1 - model includes the ScratchMaths dummy, KS1 maths attainment (pupil and school levels) and CT test score.
- Stage 2 - including ScratchMaths*CT Score interaction.

The purpose of the first model stage was to explore the impact of ScratchMaths on KS2 maths attainment in 2017 when both KS1 maths attainment (in 2013) and CT test (in 2016) are statistically controlled for. If a positive impact was found from analyses of the interim CT test, this might account for any positive impact observed in KS2 maths. In other words, the stage 1 model is exploring whether ScratchMaths had a direct impact on KS2 maths attainment once taking the interim CT test score into account. At stage 2, a ScratchMaths*CT test score interaction term was included in the model. The

purpose of doing this was to explore whether the relationship between computational thinking and KS2 mathematics attainment was stronger (or weaker) for pupils in the ScratchMaths intervention schools compared with pupils in the control schools. If the ScratchMaths*CT test score interaction terms were observed to be statistically significant, follow-on subgroup analyses would explore the impact of ScratchMaths in two subsamples: one with relatively low CT test scores (five or less out of 10) and one with relatively high CT test scores (above five).

Implementation and process evaluation

Overview

During the design year, the process evaluation aimed to provide an independent review of the process of the design of the curriculum materials and associated PD activities, and IoE's evaluation of these, and to provide guidance to the project team on ensuring that the intervention approaches, materials and training would be replicable and testable through a randomised control trial.

During the intervention years, the process evaluation aimed to evaluate the reliability and validity of any identified impact through analysis of fidelity and scalability, in particular the barriers and necessary conditions for successful implementation, and to address other evaluation research questions.

IPE data collection

IoE kept records of attendance at professional development events and undertook initial data collection on technological prerequisites and the organisation of computing in schools. Y5 participating teachers were surveyed in 2016, and Y6 participating teachers in 2017, to collect data on implementation, and a sample of these surveyed teachers were interviewed. Teachers in control schools were also surveyed. Details of samples are provided in Appendix J.

SHU and IoE teams met in November 2014 so that SHU understood IoE's plan for their curriculum design evaluation to inform the process evaluation and ensure there was no replication so that schools are not overburdened.

The following data and materials were collected from IoE to support the process evaluation:

- A report of key findings from the development work with the ScratchMaths design schools provided at the end of the design year.
- Copies of training and curriculum materials for Y5 (received prior to use with Y5) and Y6 (received prior to use with Y6).
- Information on recruited schools and teachers prior to randomisation: address, head teacher, chair of governors.
- Information before the start of intervention, where possible, about the two teachers participating in the first year of the trial and the two teachers participating in Year 2 (in most cases schools were not able to provide this data).
- Baseline information about existing use of Scratch obtained from a pre-PD survey administered by the ScratchMaths team
- Attendance records of teachers at PD events.
- Surveys on teacher views of PD events.
- Records of changes to participating teachers' activity and schools' withdrawal from participation in the PD.
- Summaries of participation in online activity.
- Information on project costs.

Table 5 below provides details of process evaluation methods²⁵.

²⁵ Details of approaches to sampling of interviewees and achieved samples of survey respondents are provided in the section below where process evaluation findings are reported and in Appendix J.

Table 5: Process evaluation methods

Pilot and design 2014/15	Review of IoE design evaluation	<p>Summary outcomes of data collected during the design phase by IoE were reviewed. Inform design of the process evaluation tools.</p> <p>Collection of data on school and teacher profiles during recruitment phase.</p>
Intervention with Y5 2015/16	<p>Visit to two Professional Development (PD) events - first and second day of training in two separate hubs.</p> <p>Telephone interviews with nine teachers in intervention schools.</p> <p>Survey of all teachers in the intervention and control schools.</p> <p>Review of IoE design evaluation data.</p>	<p>Observation of PD, informal discussion if possible with teachers and PD leaders. Key foci: fidelity in use of PD materials, the nature of the PD used.</p> <p>Semi-structured interviews focused on: experience of PD, key professional learning outcomes, use of curriculum materials, changes in practice. Key foci: fidelity in use of curriculum materials.</p> <p>Collect fidelity data on implementation in Wave 1 schools. Collect data on any other practices or activities that might influence computational thinking and/or mathematics e.g. other interventions. For intervention teachers, evaluation of PD and curriculum materials, affordances and barriers to engagement. Identify any issues that might affect balance. Also identify use of online or additional support.</p> <p>Additional data sources on fidelity inform design of evaluation tools for the following year.</p>
Intervention with Y6 2016/17	<p>Visits to two PD events (hubs not visited the previous year).</p> <p>Telephone interviews with nine teachers in intervention schools.</p> <p>Survey of all Y6 teachers in the intervention and control schools. Survey of Y5 teachers in the control schools.</p>	<p>Observation of PD, informal discussion if possible with teachers and PD leaders. Key foci: fidelity in use of PD materials and the nature of the PD used.</p> <p>Semi-structured interviews focused on: experience of PD, key professional learning outcomes, use of curriculum materials, changes in practice. Key foci: fidelity in use of curriculum materials. Identify issues of school-level professional learning community.</p> <p>Collect fidelity data on implementation in Wave 1 schools. Collect data on any other practices and activities that might influence computational thinking and/or mathematics e.g. other interventions. For intervention teachers: evaluation of PD and curriculum materials, affordances and barriers to engagement. Identify any issues that might affect balance. Also identify use of online or additional support.</p>

Interview schedules were developed by the evaluation team and reviewed by the ScratchMaths team. Surveys were co-designed with the ScratchMaths team. Summary records of interviews were made in relation to questions and analysed by open thematic coding. Survey data was used to derive descriptive statistics; further details are given in appendices J and K.

Details of the sample of telephone interviews were withheld from the ScratchMaths team and participants consented on the basis of anonymity. For teacher surveys, it was agreed with EEF that full survey data would be shared with the ScratchMaths team to facilitate their research activity and integration into their own data set. Participants were made aware of this in information sent and again at the start of the survey. This does introduce a possibility of bias in responses, given that respondents were aware the data would be shared with the ScratchMaths team.

Samples

Nine interviews were conducted with participants in 2016 (Y5 teachers) and 2017 (Y6 teachers). All teachers who participated in ScratchMaths professional development as part of either the Wave 1 Y5 or Y6 intervention sample were invited to participate in the survey, with responses from 36 schools (44 teachers) in 2016 (Y5) and 31 schools (35 teachers) in 2017 (Y6). In some cases the same teachers made more than one attempt on the same survey. The approach to analysing multiple responses from a single teacher or more than one teacher in a school in relation to implementation and fidelity is discussed below in the sub section - 'Determining levels of implementation and fidelity' and in more detail in Appendix J, which also provides further detail of the data corpus, samples and possible sample bias.

To assess the control condition Wave 2 control schools were surveyed in 2016 (Y5) and 2017 (Y6). In 2016 37 teachers from 34 schools completed the survey, thus a similar number to the intervention sample. These teachers then participated in the professional development. In 2017 there was only one response to the Wave 2 control survey; the lower number explained by the lack of email addresses for direct contact with the Y6 teachers and that they were not about to engage in ScratchMaths professional development. In addition, a further survey of Y5 teachers in the control schools in 2017 was undertaken to obtain further data to inform research questions related to teacher views on the PD and materials, as these teachers had been part of the wait list group.

Fidelity criteria

The ScratchMaths development team provided definitions for high, medium and low fidelity of the intervention in terms of attendance of PD, technology, coverage, time and progression. These criteria were not established before the trial. An initial set of criteria was developed at the end of the first full year of implementation²⁶ and then revised during the second year in March 2017. The survey for the Y6 intervention teachers was in the process of being drafted and the 2017 intervention schedule was designed after this date. Table 6 presents the revised criteria.

²⁶ Original fidelity criteria can be found for comparison in the evaluation protocol https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/Round_6-_Scratch_maths_amended.pdf

Table 6: Revised fidelity criteria for ScratchMaths intervention and data sources

	High	Medium	Low
<p>1. Professional Development</p> <p>Data from IoE PD attendance data</p> <p>Data from Y5 (Q7) & Y6 (Q9) teacher survey used for comparison purposes</p>	<p>Y5 teacher attended at least two days of PD or equivalent (defined as any combination of Summer 2015 PD days or half-day optional PD or substantial in-school PD via cascade/co-planning with a teacher who attended PD).</p> <p>Y6 teacher attended at least two days of PD or equivalent (defined as any combination of Summer 2016 PD days or half-day optional PD or substantial in-school PD via cascade/co-planning with Y5 teacher or another teacher who attended PD).</p>	<p>Y5 teacher attended at least one day of PD or equivalent (defined as any combination of Summer 2015 PD days or half-day optional PD or substantial in-school PD via cascade/co-planning with a teacher who attended PD).</p> <p>Y6 teacher attended at least one day of PD or equivalent (defined as any combination of Summer 2016 PD days or half-day optional PD or substantial in-school PD via cascade/co-planning with Y5 teacher or another teacher who attended PD).</p>	<p>Y5 teacher had some form of limited PD with a teacher who had attended PD, their SMLC or a member of the SM team.</p> <p>Y6 teacher had some form of limited PD with a teacher who attended PD or taught Y5 SM, their SMLC or a member of the SM team.</p>
<p>2. Technology</p> <p>Data from Y5 (Q14) teacher survey</p>	<p>Computers running Scratch 2.0 offline or adequate internet access</p> <p>Minimum 2:1 pupil to computer ratio²⁷</p>		<p>Computers running Scratch 2.0 offline or adequate internet access</p> <p>Minimum 3:1 pupil to computer ratio</p>
<p>3. Coverage</p> <p>Data from Y5 & Y6 teacher Surveys (15 questions relating to 3 modules in each year - 4 investigations plus 1 test per module)</p>	<p>Pupils taught at least some of the core activities across 5 different modules</p>	<p>Pupils taught at least some of the core activities from across 4 different modules</p>	<p>Pupils taught at least some of the core activities from across 3 different modules</p>
<p>4. Time</p> <p>Data from Y5 (Q13) & Y6 (Q19) teacher survey</p>	<p>Time spent on teaching is at 20+ hours in Y5 and at least 12+ hours in Y6.</p>	<p>Time spent on teaching is at least 12+ hours per year.</p>	<p>Time spent on teaching is less than 12 hours per year.</p>
<p>5. Progression</p> <p>Data from Y5 (Q13) & Y6 (Q19) teacher survey</p>	<p>The order of modules and order of activities are mostly followed in general.</p>	<p>The order of modules and order of activities are mostly followed in general.</p>	<p>The order of modules is mostly followed in general.</p>

Determining levels of implementation and fidelity

In this section, the approach to determining fidelity is considered. It was necessary to composite some of the survey data. This was due to multiple responses by a small number of teachers and also

²⁷ With regard to technology fidelity, the ratio of 2:1 was considered optimum with pupils learning collaboratively rather than 1:1.

responses by more than one teacher per school in some cases. The term 'composite' is used to refer to the process of combining individual teacher responses into a single school response. This was done by averaging data or where this was not possible or meaningful, for example, in relation to ratio of computers to children, the lower implementation level was used. Further details of fidelity data and how survey responses were analysed and composited are provided in Appendix K and J.

In the cases of attendance and implementation, fidelity was determined by, firstly, considering the overall teacher sample (ignoring the aspect of the attendance fidelity criteria related to in-school teacher-to-teacher professional development). Secondly, for attendance, fidelity is considered in relation to the Y5 (n=35) and Y6 (n=31) composite school-level survey responses. Here both the attendance criteria and the criteria related to in school professional development by an attendee were applied. For other fidelity dimensions, data are considered in relation to the composite school survey respondents to the Y5 and Y6 surveys.

Because overall fidelity is determined by combining Y5 and Y6 fidelity data across all dimensions, it is only possible to determine fidelity for the 27 cases used in the on-treatment analysis.

Data from teacher interviews suggest that where there was variance between classes in schools this was relatively minor, for example, different amounts of time spent on activities or different ways activities were introduced. Thus it is a reasonable assumption that the differences were not substantial. Therefore, for all fidelity dimensions other than attendance it is assumed that individual teacher responses apply to the whole school.

Teacher response, school response and other implementation and process evaluation analysis

Above the issue of the need to composite teacher responses to consider implementation levels and fidelity was discussed. For other aspects of the process evaluation such as teacher views of the materials or professional development outcomes or whether teachers taught ScratchMaths in mathematics or other lessons, full teacher level data is reported.

Costs

EEF cost evaluation guidance was supplied to the IoE delivery team who requested their finance department to identify costs for delivery of the intervention, after separating out costs of design, development and research by the IoE ScratchMaths team. Due to difficulties in separating costs for different aspects of the project, SHU calculated costs for models of PD delivery for delivery by two PD leads, as observed based on likely staff and non-pay costs. The method used to calculate costs followed EEF guidance on costs per pupil per year.

The delivery team was also requested to provide exemplifications of delivery costs by local hub providers based on costs of training for control schools for Y5 teachers in 2016/17 as per the waitlist design. These costs would apply to replication in the existing hubs given that the leads were already familiar with the materials and had supported the intervention trainings. If the project was scaled up, there might be additional costs to train professional development facilitators in other areas.

Timeline

The table below provides a timeline of intervention and evaluation activity.

Table 7: Timeline

Date	Activity	Responsibility
Sept 2014 - Apr 2015	ScratchMaths design and set up including work with design schools	ScratchMaths team
Aug 2014 - July 2015	Develop CT test and pilot	Evaluation team
Jan 2015 - Mar 2015	Recruit schools to trial	ScratchMaths team
May 2015	Randomisation of schools	Evaluation team
June 2015	School training begins for Y5 teachers	ScratchMaths team
July 2015	Process evaluation visits to PD event visits	Evaluation team
Sept 2015 – Apr 2016	Delivery to Y5 pupils in intervention schools	ScratchMaths team
October 2015	Design second year of materials	ScratchMaths team
Nov 2015 - Jan 2016	CT test with independent sample to establish correlation with KS2	Evaluation team
Feb - Apr 2016	Process evaluation telephone interviews with teachers	Evaluation team
May 2016	Testing pupils with CT test	Evaluation team
June 2016	Y5 teacher survey	Evaluation team
June 2016	Training for Y6 teachers begins in intervention schools. Training for Y5 Wave 2 control teachers begins	ScratchMaths team
June 2016	Process evaluation visits to PD events	Evaluation team
Sept 2016 - Apr 2017	Delivery to Y6 pupils in intervention schools and Y5 pupils in control schools	ScratchMaths
Feb 2017 - Apr 2017	Process evaluation telephone interviews with teachers	Evaluation team
May 2017 - July 2017	Survey of control and intervention school teachers	Evaluation team
Nov 2017 - Jan 2018	Retrieval of NPD data, analysis and reporting	Evaluation team

Impact evaluation

Participants

Figure 5: Participant flow diagram: primary outcome (KS2 maths attainment in Y6 in May 2017)

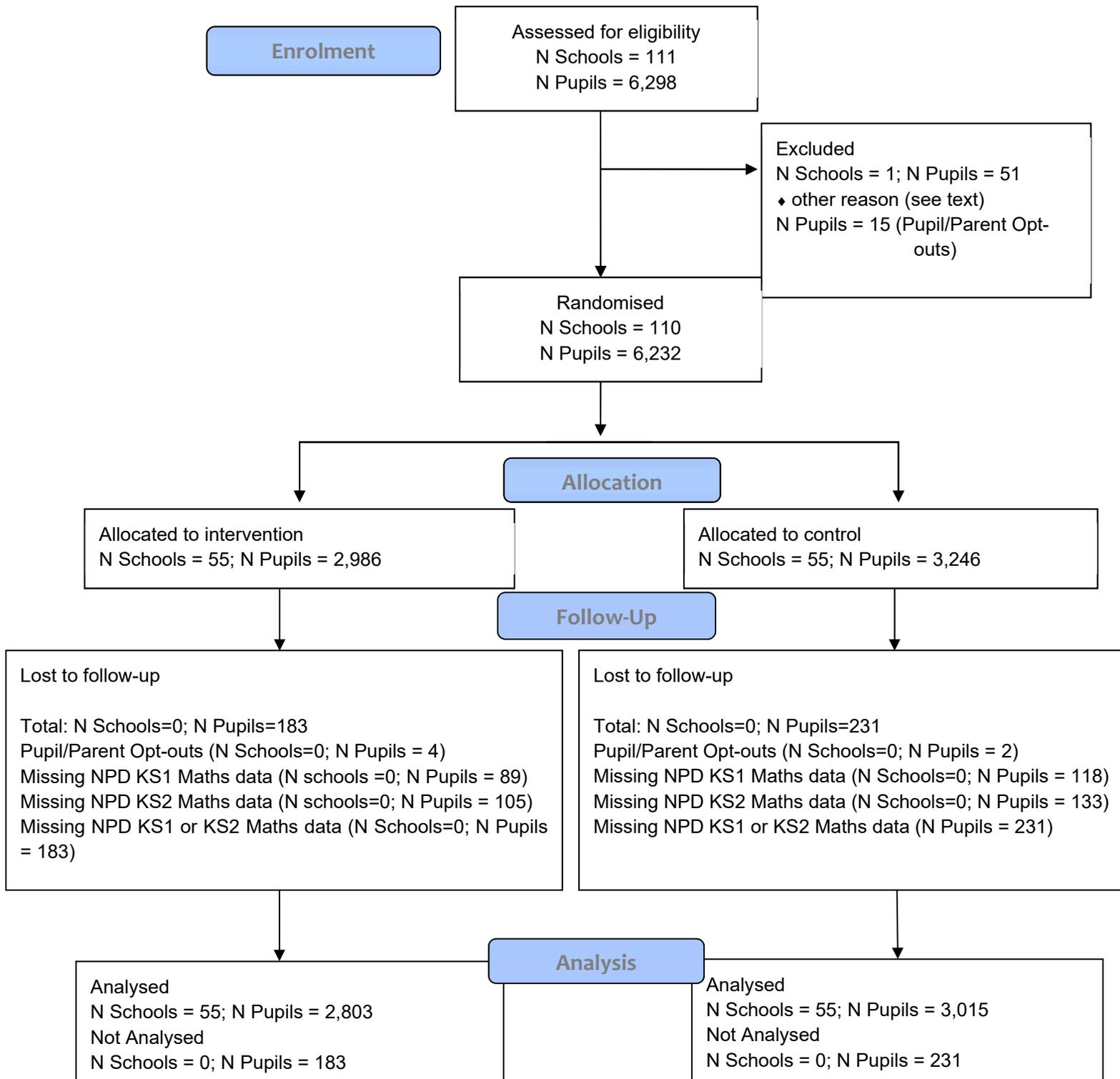


Figure 6: Participant flow diagram: interim/secondary outcome (computational thinking in Y5 in July 2016)

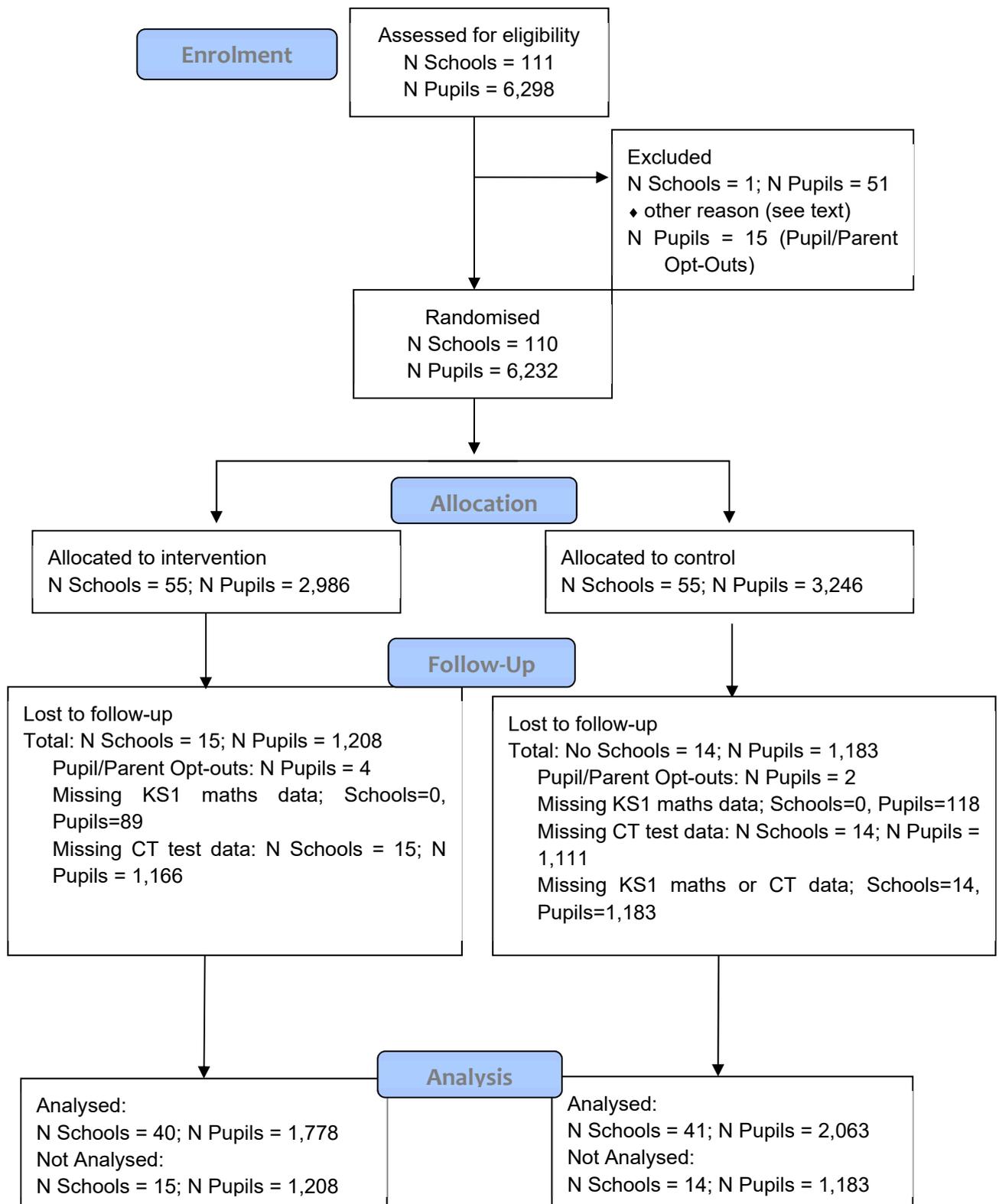


Table 8: Minimum detectable effect size at different trial stages

Stage	N [schools/pupils] (n=intervention; n=control)	Correlation between pre-test (+ other covariates) & post-test	ICC	Blocking/stratification or pair matching	Power	Alpha	Minimum detectable effect size (MDES)
Primary Outcome (KS2 Maths Attainment, 2017)							
Protocol	110 schools, 220 classes, 4,400 pupils Intervention: 55 schools, 110 classes, 2,200 pupils Control: 55 schools, 110 classes, 2,200 pupils	School 0.77	School 0.13 Class 0.07	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.18
Baseline	110 schools, 207 classes, 6,232 pupils Intervention: 55 schools, 97 classes, 2,986 pupils Control: 55 schools, 110 classes, 3,246 pupils	School 0.77	School - 0.13 Class - 0.07	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.18
Analysis (i.e. available pre- and post-test)	110 schools, 207 classes, 5,818 pupils Intervention: 55 schools, 97 classes, 2,803 pupils Control: 55 schools, 110 classes, 3,015 pupils	School 0.42 Pupil 0.71	School - 0.11 Class - 0.01	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.17
Interim Secondary Outcome (Score in Computational Thinking Test, 2016)							
Protocol	110 schools, 220 classes, 4,400 pupils Intervention: 55 schools, 110 classes, 2,200 pupils Control: 55 schools, 110 classes, 2,200 pupils	School 0.50	School 0.17 Class 0.07	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.21

Baseline	110 schools, 207 classes, 6,232 pupils Intervention: 55 schools, 97 classes, 2,986 pupils Control: 55 schools, 110 classes, 3,246 pupils	School 0.50	School 0.17 Class 0.07	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.21
Analysis (i.e. available pre- and post-test)	81 schools, 162 classes, 3,841 pupils Intervention: 40 schools, 74 classes, 1,778 pupils Control: 41 schools, 88 classes, 2,063 pupils	School 0.60 Pupil 0.49	School 0.13 Class 0.02	Pair matching using propensity scores, blocked by geographical area	80%	0.05	0.18

Table 8 shows that for the primary and interim secondary outcomes, the statistical precision of the trial was better at the analysis stage than was estimated in the protocol or at baseline (as shown by the smaller minimum detectable effect size (MDES) estimates at the analysis stage). The small improvement in estimated precision can be accounted for in two ways. First, the statistical strength of clustering at the school and class levels for the two outcome variables was slightly lower than originally estimated. This is shown in Table 8 by smaller school-level and class-level Intra Cluster Correlation coefficients (ICCs) at the analysis stage. Second, at the analysis stage we included explanatory power of KS1 maths attainment at both school and pupil levels. Whilst the stronger explanatory power of KS1 maths and weaker levels of school and class-level clustering account for the smaller MDES estimates shown in Table 8, some caution is advised. This is because the smaller MDES estimates at the analysis stage are drawing on data with some missing values. This is less of an issue with the primary KS2 maths outcome (414 cases missing, 6.6% of all baseline cases) than for the interim CT test (2,391 cases missing, 38.4%). Missing data for both the primary and interim secondary outcomes are explored for patterns within the impact analyses below and taken into account through sensitivity analyses discussed below.

Pupil characteristics

Table 9 below compares the intervention and control school baseline samples, showing an excellent balance at both school and pupil levels across all measures except OFSTED inspections and the percentage of pupils with English as an additional language (EAL). In terms of the pre-test KS1 maths covariate, the difference between the intervention and control samples has an effect size of 0.03 sds. In terms of OFSTED inspections: intervention schools were less likely to be classed as 'outstanding' (9 schools, 17%) compared with control schools (18 schools, 35%) and more likely to be classed as 'requires improvement' (17% of intervention schools compared with 4% of control schools). The proportion of EAL pupils was slightly higher in controls schools (30%) compared with intervention schools (24%).

Table 9 Baseline comparisons

Variable	Intervention group		Control group	
School-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
OFSTED Inspection (Overall Effectiveness - Most recent inspection (at end of 2015) :	N=53 (2)		N=52 (3)	
1 - Outstanding	9 / 53	17.0%	18 / 52	34.6%
2 - Good	35 / 53	66.0%	32 / 52	61.5%
3 - Requires Improvement	9 / 53	17.0%	2 / 52	3.8%
Type of schools:	N=55		N=55	
Sponsored Academy	0(0)	0.0%	3(0)	5.5%
Converter Academy	4(0)	7.3%	3(0)	5.5%
Community Foundation	32(0)	58.2%	32(0)	58.2%
Voluntary Aided	2 (0)	3.6%	5 (0)	9.1%
Voluntary Controlled	11 (0)	20.0%	7 (0)	12.7%
	6 (0)	10.9%	5 (0)	9.1%
School-level (continuous)	n (missing)	Mean	n (missing)	Mean
Data from School Census (2014 academic year)				
2014 KS1 Average Points Score	55 (0)	14.9	55 (0)	14.9
2014 KS1 to KS2 Value Added (maths)	55 (0)	100.4	55 (0)	100.3
2014 % with level 5+ in KS2 Maths	55 (0)	42.3	55 (0)	41.8
% FSM in last 6 years	55 (0)	32.5	55 (0)	31.8
% EAL	55 (0)	24.1	55 (0)	30.2
% SEN (Statement or School Action)	55 (0)	11.5	55 (0)	11.3
% Female	55 (0)	48.9	55 (0)	49.4
School Size	55 (0)	397	55 (0)	426
Pupil-level NPD data aggregated to the school level				
Aggregated KS1 Maths Point Score	55 (0)	16.0	55 (0)	16.0
Aggregated KS1 Average Point Score	55 (0)	15.7	55 (0)	15.7
Pupil-level (categorical)	n/N (missing)	Percentage	n/N (missing)	Percentage
Ever eligible for free school meals (EVERFSM_ALL_SPR13)	830/2,985 (91)	28.7%	885/3,128 (118)	28.3%
Female	1,436 / 2,898 (88)	49.6%	1,560 / 3,130 (116)	49.8%
Pupil-level (continuous)	n (missing)	[Mean or median]	n (missing)	[Mean or median]
KS1 Average Points Score	2,897 (89)	15.8	3,128 (118)	15.7
KS1 Maths Points Score	2,897 (89)	16.1	3,128 (118)	16.0

Missing data

The quantity of missing data was much lower within the impact analyses for the primary KS2 maths outcome (414 cases missing, 6.6% of all baseline cases) compared with the interim CT test (2,391 cases missing, 38.4%). The impact of missing data on the balance between the intervention and control group samples is illustrated descriptively in Table 10 in terms of FSM status, gender and KS1 maths attainment. Statistics from four samples are shown for each of these factors. First, at the top, statistics

for the full sample at baseline are shown. Below this, statistics from the sample included in the main ITT impact analyses for the primary outcome (KS2 maths) are shown. Below this, statistics based on two samples that relate to the impact analyses for the interim CT test outcome are shown: first, statistics from the raw sample of pupils in the complete sample of 40 intervention schools and 41 control schools with CT test data; second, statistics from the 'complete pairs' restricted subsample of pupils in the 31 intervention schools and their 31 matched control schools with CT test data.

Table 10: Impact of missing data on the balance of intervention and control group samples for KS2 maths and CT test analyses

	Intervention group		Control group	
	n/N (missing)	Percentage	n/N (missing)	Percentage
<i>% Ever Classed as FSM [EVERFSM_ALL]</i>				
Baseline (N=6,232)	830/2,895 (91)	28.7%	885/3,128 (118)	28.3%
Primary Outcome ITT Analysis (N=5,818)	788/2,800 (3)	28.1%	844/3,013 (2)	28.0%
CT Test (Raw sample, N=3,841)	517/1,777 (1)	29.1%	595/2,062 (1)	28.9%
CT Test (Paired sample, N=3,077)	435/1,446 (1)	30.1%	522/1,629 (1)	32.0%
<i>Gender (% Female)</i>				
Baseline (N=6,232)	1,436 / 2,898 (88)	49.6%	1,560 / 3,130 (116)	49.8%
Primary ITT Analysis (N=5,818)	1,401 / 2,803 (0)	50.0%	1,518 / 3,015 (0)	50.3%
CT Test (Raw sample, N=3,841)	885 / 1,778 (0)	49.8%	1,055 / 2,063 (0)	51.1%
CT Test (Paired sample, N=3,077)	717 / 1,447 (0)	49.6%	833 / 1,630 (0)	51.1%
Pupil-level (continuous)	n (missing)	Mean (sd)	n (missing)	Mean (sd)
<i>KS1 Maths Points Score</i>				
Baseline (N=6,232)	2,897 (89)	16.1 (3.44)	3,128 (118)	16.0 (3.44)
Primary ITT Analysis (N=5,818)	2,803 (0)	16.2 (3.35)	3,015 (0)	16.2 (3.25)
CT Test (Raw sample, N=3,841)	1,778 (0)	16.0 (3.45)	2,063 (0)	16.2 (3.40)
CT Test (Paired sample, N=3,077)	1,447 (0)	16.0 (3.46)	1,630 (0)	16.0 (3.43)

As reported above, at baseline the difference between the intervention and control group sample in terms of KS1 attainment was small (an effect size of +0.03 sds). For the primary ITT analyses which exclude the 414 pupils with missing KS2 or KS1 data, the difference is zero. For the analyses of the interim CT test outcome, within the raw sample, the difference was small but larger than at baseline and in a different direction (-0.06 sds). Follow-on sensitivity analyses restricted the sample to complete pairs of intervention and control schools with CT test data. When this was done, the difference returned to zero. As discussed in the randomisation section above, this is an illustration of how the propensity-score-paired-stratification design is robust to whole schools drop outs (15 intervention and 14 control schools here). Specifically, this illustrates how this design can be used to best ensure good balance (albeit with a reduced sample and hence statistical power).

After examining the impact of missing values on the baseline balance²⁸, we feel confident that our research design and analysis plan was robust enough to be confident of our findings from the impact analyses for the primary outcome (overall maths attainment) and follow-on secondary outcomes (attainment in the three KS2 maths test papers). For the primary outcome, missing values were looked

²⁸ For the primary outcome, this is illustrated in Table 9 by comparing statistics for the baseline and primary outcome ITT analyses. For the secondary outcome it is illustrated in Table 9 by comparing statistics for the baseline and CT test raw sample analyses. The re-balancing provided by the propensity-score-paired-stratification for the CT test secondary outcome is illustrated in Table 9 by comparing the baseline, raw and paired statistics.

at directly and found to be weakly correlated with KS1 attainment in maths ($r=-0.21$) and overall (-0.22), were more likely to be male (4.4%) compared with female (2.6%) and more likely to have been classed as FSM (4.8%) compared with pupils not classed as FSM (2.9%). These patterns were consistent for both intervention and control group samples which is reflected by the excellent balance shown in Table 9 at baseline and for the primary outcome ITT analysis

The missing data for the interim CT test outcome are more problematic and meant that an ITT approach for the impact analyses was precluded. The patterns in Table 10 suggest that the planned complete pairs sensitivity analyses will help to ensure a good balance in terms of KS1 mathematics between intervention and control samples. Whilst this does not completely eliminate the risk of bias brought by missing data, we feel that this approach provides a useful way of scrutinising the impact analysis finding.

Descriptive summary

Prior to presenting the multilevel impact analyses, Table 11 presents a descriptive summary of the primary and secondary outcomes for the intervention and control group samples in the ScratchMaths evaluation. From this table, the largest impact for ScratchMaths is observed to be with the interim CT test and this is relatively small (effect size = $+0.10$ sds). For the primary KS2 maths outcome and across the three KS2 maths test papers, the impact is observed to be close to zero.

It would not be appropriate to use the descriptive summary to determine whether ScratchMaths had a causal impact on KS2 maths attainment. This is because the statistics presented in Table 10 do not take account of how pupils are clustered into schools within geographical areas and into classes within schools, nor do they control for different levels of KS1 maths attainment. However, area/school and class clustering and KS1 maths attainment are both taken into account within the multilevel analyses used to evaluate the causal impact of ScratchMaths that are presented in the next section.

Table 11: Descriptive summary of ScratchMaths outcome variables

	Intervention group		Control group		E.S.
Overall KS2 Maths Attainment	n (missing)	Mean (sd.)	n (missing)	Mean (sd.)	Hedges g
KS2 Maths (Raw Points)¹	2,877 (105)	76.2 (23.85)	3,111 (133)	76.5 (23.46)	-0.01
KS2 Maths (Scaled)	2,877 (105)	104.9 (7.26)	3,108 (136)	105.0 (7.09)	-0.01
KS2 Maths Test Papers	n (missing)	Mean (sd.)	n (missing)	Mean (sd.)	
KS2 Maths Paper 1 (Arithmetic)	2,877 (105)	31.4 (8.08)	3,112 (132)	31.8 (7.87)	-0.05
KS2 Maths Paper 2 (Reasoning 1)	2,878 (104)	23.8 (8.54)	3,112 (132)	23.9 (8.85)	0.00
KS2 Maths Paper 3 (Reasoning 2)	2,879 (103)	21.0 (8.77)	3,111 (133)	20.9 (8.68)	+0.01
Computational Thinking	n (missing)	Mean (sd.)	n (missing)	Mean (sd.)	
CT Test Score (Raw)	1,820 (1,162)	4.95 (2.22)	2,136 (1,108)	4.73 (2.18)	+0.10
CT Test Score (Complete Pairs)	1,483 (1,502)	4.85 (2.21)	1,688 (1,483)	4.64 (2.20)	+0.10

Note. These are bivariate statistics and so have fewer missing values (238 for the raw KS2 maths primary outcome) compared with the multivariate ITT analysis (414 missing cases, see Figure 5). The supplementary, scaled KS2 maths measure had slightly more missing cases (241).

Impact analyses for primary outcome

Table 12 summarises the headline ITT analyses for the KS2 maths attainment primary outcome obtained from the model that included KS1 maths attainment... This shows that we found no evidence of impact for ScratchMaths on the primary outcome (Hedges g effect size = 0.00). This finding was consistent for models using either the raw or scaled KS2 maths attainment variables.

Table 12: Summary of headline multilevel ITT impact analyses for primary outcome (KS2 maths)

	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
Raw KS2 Maths Attainment	2,877 (105)	76.2 (75.3; 77.1)	3,111 (133)	76.5 (75.7; 77.3)	5,818 (2,803; 3,015)	0.00 (-0.12; +0.12)	0.970
Scaled KS2 Maths Attainment	2,877 (105)	104.9 (104.6; 105.2)	3,108 (136)	105.0 (104.8; 105.2)	5,815 (2,803; 3,012)	+0.01 (-0.11; +0.12)	0.933

Appendix I provides additional details on the multilevel models used to estimate the impact of ScratchMaths on overall KS2 maths attainment. It also includes specific details on how the above Hedges g effect size was calculated from the model, school and class ICC statistics and explanatory power for KS1 maths at school and pupil levels.

No evidence was found that the impact of ScratchMaths on KS2 maths attainment interacted with FSM status, gender or KS1 maths attainment. Table 13 summarises the ITT analyses for the KS2 maths attainment primary outcome for FSM and not-FSM pupil subsamples. We found no evidence for ScratchMaths having an impact on KS2 maths attainment for either subsample (Hedges g effect size = +0.01 for both). This finding was consistent for models using either the raw or scaled KS2 maths attainment variables.

Table 13: Summary of headline multilevel ITT impact analyses for primary outcome (KS2 maths) by FSM

	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
FSM subsample	789 (41)	70.7 (69.0; 72.4)	844 (41)	69.4 (67.8; 71.0)	1,632 (788; 844)	+0.01 (-0.14; +0.16)	0.915
not-FSM subsample	2,012 (53)	78.5 (77.5; 79.5)	2,170 (73)	79.5 (78.6; 80.4)	4,181 (2,012; 2,169)	+0.01 (-0.11; +0.13)	0.874

Impact analyses for follow-on secondary outcomes (KS2 maths test papers)

Table 14 summarises the ITT analyses for the three separate KS2 maths tests sat by participants in 2017. As shown in Table 14, we found no evidence for ScratchMaths having an impact on pupil attainment on any of the separate KS2 maths tests.

Table 14: Summary of multilevel ITT impact analyses for follow-on secondary outcomes (KS2 maths test papers)

Outcome	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention ; control)	Hedges g (95% CI)	p-value
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
Paper 1 - Arithmetic	2,877 (105)	31.4 (31.1; 31.7)	3,112 (132)	31.8 (31.5; 32.1)	5,819 (2,803; 3,016)	-0.04 (-0.16; +0.08)	0.542
Paper 2 - Reasoning 1	2,878 (104)	23.8 (23.5; 24.1)	3,112 (132)	23.9 (23.6; 24.2)	5,820 (2,804; 3,016)	+0.01 (-0.11; +0.13)	0.859
Paper 3 - Reasoning 2	2,879 (103)	21.0 (20.7; 21.3)	3,111 (133)	20.9 (20.6; 21.2)	5,820 (2,805; 3,015)	+0.03 (-0.10; +0.16)	0.637

Impact analyses for interim computational thinking secondary outcomes (2016)

Appendix H provides some descriptive detail on the distribution of the interim CT test outcomes.

Table 15 summarises the analyses for the interim CT test sat by participants in 2016. Two analyses are summarised. The first is based on the raw sample of 81 schools where CT test data were collected. The second is based on a restricted 'complete pairs' subsample of 62 of these 81 schools. The complete pairs subsample is the 31 intervention and 31 matched control schools where CT data are available for both.

Table 15 shows that ScratchMaths had a statistically significant positive effect on CT test scores (Hedges $g=+0.15$ sds). Within the complete pairs subsample analyses, the impact remains positive but is weaker and does not reach statistical significance.

Whilst statistically significant, the +0.15 effect size is below the analysis stage 0.17 MDES estimate (see Table 8), and so the statistical power will be lower than 80%²⁹. A reduction in power indicates an increased chance of false-positive (concluding an effect when one does not exist). Using the data in Table 8 a retrospective power of 60% can be estimated. The sizeable problem of missing data further increases the need for caution in interpreting impact from these models. Finally, the complete pairs sensitivity analyses resulted in a smaller effect size estimate (+0.12 sds) which was not statistically significant.

²⁹ The 0.15 effect size has an estimated statistical power of 60%.

Table 15: Summary of multilevel impact analyses for interim CT test secondary outcomes

Outcome	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
Raw (All Participants)	1,820 (1,162)	4.95 (4.85; 5.05)	2,136 (1,108)	4.73 (4.64; 4.82)	3,841 (1,778; 2,063)	+0.15 (+0.001; +0.29)	0.048*
Complete Pairs subsample	1,4883 (1,503)	4.85 (4.74; 4.96)	1,688 (1,483)	4.64 (4.54; 4.75)	3,077 (1,447; 1,630)	+0.12 (-0.04; +0.29)	0.142

*p<0.05

We found no evidence that the impact of ScratchMaths on CT test score interacted with gender or KS1 maths attainment. However, a statistically significant interaction was found with FSM status.

Table 16 summarises the ITT analyses for the CT test outcomes for FSM and not-FSM pupil subsamples. Behind the cautious 'positive impact' found for all pupils, we found notable differences relating to FSM status. For FSM pupils, the impact of ScratchMaths on CT test scores was positive, larger and statistically significant ($g=+0.25$). For the not-FSM subsample the effect size was smaller and not statistically significant ($g=+0.10$). Given that these are subsample analyses that this trial was not designed or powered to detect along with the missing data, the findings must be considered as tentative. However, our analyses do suggest that ScratchMaths had a moderate positive impact on CT test scores at the end of Y5, half-way through the trial.

Table 16: Summary of multilevel impact analyses for interim CT test secondary outcomes by FSM

Outcome	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)				
FSM subsample							
Raw	517 (313)	4.61 (4.42; 4.80)	595 (290)	4.13 (3.96; 4.30)	1,112 (517; 595)	+0.25 (+0.08; +0.42)	0.004
Complete-pairs	435 (395)	4.57 (4.36; 4.78)	522 (363)	4.11 (3.93; 4.29)	957 (435; 522)	+0.24 (+0.04; +0.43)	0.017
Not-FSM subsample							
Raw	1,260 (805)	5.11 (4.99; 5.23)	1,468 (775)	5.02 (4.91; 5.13)	2,727 (1,260; 1,467)	+0.10 (-0.05; +0.25)	0.172
Complete-pairs	1,011 (1,054)	5.00 (4.86; 5.14)	1,108 (1,135)	4.93 (4.80; 5.06)	2,118 (1,011; 1,107)	+0.08 (-0.09; +0.24)	0.361

On-treatment analyses for primary outcome

As outlined in Appendix K, fidelity to ScratchMaths was measured at the school level and drew on five fidelity dimensions: PD attendance, IT provision, use of ScratchMaths module materials, ScratchMaths

curriculum time, and the order/progression of ScratchMaths module. The fidelity analyses were restricted to a subsample of 27 of the 55 ScratchMaths intervention schools where IPE teacher survey data was obtained for both Y5 and Y6 surveys. As is discussed in Appendix J, apart from attendance data which were provided by the ScratchMaths team at teacher level, implementation data were gathered at school level and assumptions made about the applicability of teacher response to all classes in each school.

Among these 27 schools, five were identified as having high fidelity to ScratchMaths over the two-year trial period; attendance of at least two PD days in both Y5 and Y6; at least 2:1 ratio of pupils to computers; Taught all 3 ScratchMaths modules in Y5 and at least 2 of the 3 modules in Y6; spent at least 20+ hours teaching ScratchMaths in Y5 and at least 12 hours in Y6 and followed the specified module order.

A further eight schools were identified as having medium fidelity which resulted in a sample of 13 identified as having medium or high fidelity to ScratchMaths. attendance of at least one PD days in both Y5 and Y6; at least 2:1 ratio of pupils to computers; Taught at least 2 of the 3 ScratchMaths modules in both Y5 and Y6; spent at least 12+ hours teaching ScratchMaths in both Y5 and Y6 and followed the specified module order.

Table 17 summarises the on-treatment analyses exploring evidence of impact for ScratchMaths on KS2 maths attainment for the subsample of pupils located in one of the five high-fidelity or 13 medium/high-fidelity intervention schools.

Attainment in KS2 maths for pupils in the restricted fidelity intervention school subsamples was compared with the attainment for pupils in control schools. First, this was done using the raw control sample of 55 schools as the comparison group. The second analysis drew on the propensity-score-paired-school-stratification research design to limit the control sample to include only control schools that were paired with the five high-fidelity or 13 medium/high-fidelity intervention schools prior to randomisation.

Table 17 Summary of multilevel on-treatment impact analyses for primary outcome (KS2 maths)

Outcome	Bivariate Descriptive Statistics				Effect size from stage 2 multilevel multivariate analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
High Fidelity							
Raw	277 (9)	77.8 (75.1;80.5)	3,111 (133)	76.5 (75.7;77.4)	3,287 (272; 3,015)	+0.01 (-0.29; +0.32)	0.919
Complete Pairs	277 (9)	77.8 (75.1;80.5)	323 (13)	83.1 (80.9;85.2)	584 (272; 312)	-0.18 (-0.42; +0.07)	0.152
Medium or High Fidelity							
Raw	664 (25)	71.8 (69.9;73.6)	3,111 (133)	76.5 (75.7;77.4)	3,665 (650; 3,015)	-0.13 (-0.33; +0.07)	0.190
Complete Pairs	664 (25)	71.8 (69.9;73.6)	721 (37)	76.9 (75.3;78.6)	1,349 (650; 699)	-0.05 (-0.23; +0.14)	0.617

Among schools identified as having high or medium/high fidelity to the ScratchMaths intervention across the two years of the trial, we found no evidence that ScratchMaths had a positive impact on KS2 maths attainment.

Follow-on exploratory analyses for primary outcome

RQ3 for this evaluation was to examine the correlation between computational thinking and mathematics attainment. Table 18 summarises the bivariate correlation coefficients between KS1 maths (in 2013), CT test score (2016) and KS2 maths (2017).

Table 18: Pupil-level Pearson Correlation Coefficients between KS1 maths, CT test score and KS2 maths for intervention and control group samples

	Intervention Group			Control Group		
	KS1 Maths (2013)	CT Test (2016)	KS2 Maths (2017)	KS1 Maths (2013)	CT Test (2016)	KS2 Maths (2017)
KS1 Maths (2013)	1.00 n=2,897			1.00 n=3,128		
CT Test (2016)	+0.49 n=1,778	1.00 n=1,820		+0.49 n=2,063	1.00 n=2,136	
KS2 Maths (2017)	+0.69 n=2,803	+0.41 n=1,779	1.00 n=2,877	+0.67 n=3,015	+0.50 n=2,081	1.00 n=3,111

The correlation between KS1 maths attainment and CT test score is of a similar magnitude for both intervention and control groups (both $r=+0.49$). The correlation between CT test score and KS2 maths attainment is smaller for the intervention group ($r=+0.41$) compared with the control group ($r=+0.50$). The correlation between KS1 and KS2 maths attainment is similar for the intervention ($r=+0.69$) and control group ($r=+0.67$) pupil samples.

Table 19 summarises the follow-on multilevel analyses exploring the impact of ScratchMaths on KS2 maths attainment whilst statistically controlling for both CT test score and KS1 maths attainment.

We found no evidence that ScratchMaths had a positive impact on KS2 maths attainment when KS1 maths attainment and CT test score were controlled for. Further, we found no evidence to suggest the relationship between computational thinking (as measured by the CT test) and KS2 maths attainment differed for pupils in the ScratchMaths intervention schools compared with pupils in control schools (as shown by the very small and not significant ScratchMaths*CT test interaction in Table 19).

Table 19: Summary of exploratory analyses for primary outcome (KS2 maths)

Outcome	Raw means				Effect size from stage 2 multilevel analyses		
	Intervention group		Control group		n in model (intervention; control)	Hedges g (95% CI)	p-value
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)			
From Main Analyses - Primary Outcome							
Raw KS2 maths attainment, controlling for KS1 maths attainment							
All Participants	2,877 (105)	76.2 (75.3; 77.1)	3,111 (133)	76.5 (75.7; 77.3)	5,818 (2,803; 3,015)	0.00 (-0.12; +0.12)	0.970
Exploratory Follow-on Analyses - Primary Outcome							
Raw KS2 maths attainment, controlling for KS1 maths attainment & CT test score [stage 2 main effects model]							
All Participants	1,779 (41)	76.1 (75.0;77.1)	2,081 (55)	76.6 (75.6;77.6)	3,758 (1,739; 2,019)		
					ScratchMaths	-0.02 (-0.17; +0.14)	0.847
					ScratchMaths	-0.01 (-0.16; +0.14)	0.898
					ScratchMaths* CT test	-0.01 (-0.04; +0.01)	0.253
All Participants (including ScratchMaths*CT test interaction)							

In response to (cautiously) finding a significant positive impact for the interim CT test outcome that differed for FSM and not-FSM pupil subsamples, follow-on subsample exploratory analyses were undertaken. Specifically, follow-on analyses explored the impact of ScratchMaths on KS2 maths attainment while controlling for KS1 maths attainment and CT test score separately for FSM and not-FSM pupil subsamples. We found no evidence that ScratchMaths had a positive impact on KS2 maths attainment when KS1 maths attainment and CT test score was controlled for either FSM or not-FSM pupil subsamples.

Costs

Calculating costs

The development, delivery and IoE research activity during the ScratchMaths project was integrated and so not monitored separately. To provide information for the ScratchMaths team design evaluation, at PD events there were frequently more members of the team than needed to lead PD.

At least two members of the ScratchMaths team were present at all PD sessions led by them. To calculate costs, estimates have been made of likely costs of delivery separate from design, development and research activity, for delivery by two PD leads.

All costs for PD were covered by EEF and so there were no costs to schools (other than any supply cover costs if they were required).

ScratchMaths as delivered

Delivery costs are based on all 55 schools allocated to the intervention condition and the 2986 pupils identified as participating at the time of randomisation. Costs are calculated regardless of actual participation.

Schools attended four days' PD across two academic years and could access three optional twilight sessions and an online webinar. Table 20 provides summary costs for ScratchMaths team, including; staff pay costs, payment to local hosts (including venue hire, refreshments and staff time for support), travel and delivery team accommodation (where applicable). Although programme activity took place over two years (programme costs) costs are also provided as costs per year per school and pupil over three years.

Table 20: Summary costs of ScratchMaths delivery

	Programme cost	Cost per year (over 3 years)
Per school	£1,843	£614
Per pupil	£34	£11

The costs per professional development day were £154 (two PD leads). These are comparable to commercial training courses for teachers.

It should also be noted that one interviewee observed that they had saved printing costs because of the format of ScratchMaths materials in comparison with a commercial scheme used previously that required printing of a large number of worksheets for pupils to complete. However, potential printing savings are not included, as cost savings would depend on which alternative approach had been implemented by a school and replaced by ScratchMaths.

Exemplar costings for future professional development

In addition to providing information on costs for the intervention as delivered, the ScratchMaths team also provided estimates of potential costs if delivered in the future by ScratchMaths local coordinators.

Table 21, below, shows exemplar average budgets if delivered by local hubs with the same pattern of PD - that is, a total of six days' equivalent PD over two years. This is based on averaging potential indicative costs for Merseyside, London and Somerset

Table 21: Exemplar budgets for future PD

Item	Average hub costs
Venue hire and catering	£1,698
Administration	£233
Tutor planning and delivery	£3,305
Printing materials	£103
TOTAL	£5,340

The table below provides a comparison for three hubs for future PD delivery based on the same average hub and class size as in the delivered ScratchMaths project (eight schools per hub, two teachers per school, two classes and 27 pupils per class). Costs could be lower if more schools attended.

Table 22: Costs for future local delivery of ScratchMaths

Costs	
Per school (three years)	£455
Per pupil (three years)	£8

Costs for local delivery are lower than in the trial because of staff costs savings due to less time needed for travel for PD leads and other costs. However, costs would likely be a little higher for new hubs that had previously not been involved in ScratchMaths as there would potentially be 'train the trainer' professional development needed, led by the ScratchMaths team.

Costs per pupil, above, would apply when staffing in schools is stable and ScratchMaths is embedded into Y5 and Y6 curriculum (maths or computing)

Future costs also do not include any further development to materials or any maintenance further development of the ScratchMaths website, which currently is maintained as a legacy of the trial, or its maintenance costs. Delivery costs do not include costs to schools for attendance at professional development events. Assuming cover is not needed for optional twilight events, a total of two days cover are needed per teacher per year. If schools have two teachers participate per year as per the ScratchMaths professional development programme design, a total of eight days cover would be needed.

Process evaluation

Process evaluation methods were outlined above, as were details about samples; in addition, further details are provided in Appendix J about the survey and interview samples which were key data sources for many of the findings reported in this section. For convenience key information about the interview and survey samples are:

- Nine telephone interviews were conducted in 2016 with Y5 teachers, and the same number in 2017 with Y6 teachers.
- All Y5 Wave 1 intervention teachers were invited to complete a survey in 2016, as were Y5 Wave 2 control teachers.
- All Y6 Wave 1 intervention teachers were invited to complete a survey in 2017, as were both Y5 and Y6 Wave 2 control teachers.

Although teachers were asked about professional development attendance in the survey, data gathered by the ScratchMaths team at PD events was used for fidelity analysis as it was more reliable.

Issues of potential sample bias in interview and survey responses to highlight are:

- Only one of the 15 schools that were not classified as having sustained participation (see below) was represented in the survey and none in the interview sample.
- Further, analysis of reported rates of attendance in schools completing the survey indicates a higher level of attendance fidelity than across the whole sample and this is particularly marked in the case of respondents to the Y6 survey.

Appendix L includes supplementary data on the satisfaction rates collected by the delivery team after each PD event

Implementation

In this section, the following aspects of implementation are considered: school-level implementation; professional development implementation; by whom, and when, was ScratchMaths taught; and use of materials and teacher mediation. Data sources and analysis are discussed in Appendix J. In summary, where survey data are used, school-level composite data are reported about implementation of ScratchMaths for 36 Y5 schools and 31 Y6 schools. The approach to compositing data to determine a school level of implementation is provided in Appendix J. When reporting teacher views, beliefs and practices, teacher-level data are used.

School-level implementation

Of the 55 Wave 1 schools allocated to the intervention condition at randomisation, one school did not send anyone to the initial PD, while 10 schools informed the ScratchMaths team at some point during the project (generally before the second PD year or shortly after) that they were no longer using ScratchMaths materials nor intended to attend professional development (see page 59 for reasons given by schools). A further five schools indicated that they might stop using the materials or attending. However, their use of materials in Y6 is not known in most cases as only one school completed the survey. Y6 attendance from these five schools was low or did not happen. Thus, 39 of 55 schools can be considered as having sustained participation until the end of the trial (sustained participation does not equate with meeting fidelity criteria but schools in this category had the potential to do so). In relation to the trial, given that these schools did not withdraw from the trial by informing SHU as evaluators, data from these schools are included in the intention-to-treat analysis and, where survey data were already obtained from the schools, were included in implementation and process evaluation analysis.

Professional development implementation

Professional development was implemented largely as planned (see intervention description in the introduction). One minor change was to offer webinar support rather than a second twilight session in the second year for Y6 teachers. This change was informed by teacher feedback on the first year PD. Additionally, in one hub, local hub leads delivered catch-up sessions for some teachers who had missed the summer professional development. In another hub, in the second year, the hub lead led school-based professional development in one school.

Attendance data

A total of 170 teachers attended at least one PD event in either Y5 (105 teachers) or Y6 (65) and of these 25 attended in both Y5 and Y6. The intention was that 2 teachers per school (110) per year would attend PD events.

Table 23, below, provides details of PD attendance by teachers who attended at least one PD event in terms of the amount of PD attended. Summer PD days counted as one day and twilights as 0.5 days. Participation in webinars is not included because teacher-level data were not available. There was some additional participation in PD led by hub leads. The 'All schools' columns provide details of teachers from any school that were assigned to the intervention condition; this included one school that did not engage from the start of the project. As defined above, sustained participating schools are those schools that had not indicated they would or might withdraw.

Table 23: Frequency of PD attendance by participating teachers

PD (days)	All schools: number of days teachers participated		Sustained participating schools: number of days teachers participated	
	Y5	Y6	Y5	Y6
0.5	4	12	3	11
1	24	10	14	7
1.5	4	2	3	2
2	36	23	28	19
2.5	15	18	13	18
3	22	N/A	18	N/A
Total	105	65	79	57
Percentage of 110 intended teacher participants	95.5	59.1	71.8	51.8

Table 23 shows that fewer teachers attended at least one professional development event in Y6 than Y5. However this lower attendance is only partly explained by school-level withdrawal. This can be deduced because when we examine only those schools that demonstrated sustained participation, there is still a decrease from 79 individual teachers in sustained participating schools attending Y5 PD to 57 attending Y6 PD. Thus, the average number of teachers per school attending decreased from Y5 to Y6, irrespective of whether their participation in the programme was sustained or not over the two years of implementation.

Considering change in total attendance in each school, the *number of teachers* attending:

- decreased in 37 schools
- stayed the same in 14 schools
- increased in 6 schools

However, with regard to the 6 schools in which the number of teachers attending increased, this did not necessarily mean an increase in the total number of PD days attended overall per school. Also there was no guarantee that attendance in Y6 built on previous PD in Y5. The mean attendance per teacher, for those who attended at least one PD event, as a percentage of total possible attendance, was nearly identical - being 69% for both Y5 and Y6. In summary:

- From Y5 to Y6, there was a decrease in number of schools that had teachers attend any ScratchMaths PD (from 54 schools to 44 schools)
- For those schools that did have teachers attend PD in Y6, on average fewer teachers attended than in Y5
- Considering the number of days attendance per teacher attending, then for teachers who did attend PD events, a similar number of mean days were attended in Y5 (1.97 days) and Y6 (1.69 days); and when considered as a proportion of possible attendance (3 days maximum in Y5 and 2 days maximum in Y6) then the mean proportional attendance is almost the same (66% in Y5, 68% in Y6)

Details of professional development attendance are discussed further below in the section on fidelity and in Appendix J.

By whom and when was ScratchMaths taught

The intention was that ScratchMaths would be taught by the Y5 teachers and then the Y6 teachers. Whilst this question was not asked of those who completed the survey, the nine interviewees who took part in the IPE were asked who taught ScratchMaths. In Y5, eight instructors interviewed were the regular class teacher and one was a higher level teaching assistant (HLTA). Of the nine Y6 interviewees three interviewees reported that it was the computing teacher in their schools, five interviewees reported it was the regular class teacher in their schools and in one interview reported it was an HLTA.

In the survey, a similar pattern is found by considering the roles declared by respondents and their responses on classes taught. Data is provided here in relation to Y6 because of the greater mathematical focus in Y6 materials and where the issue of whether the ScratchMaths team also taught mathematics may be of greater importance. In the Y6 survey, ScratchMaths was taught by five teaching assistants (from a sample of 35 teachers), usually by the class teacher but in some cases by the computing teacher where there was some division of teaching within Y6 teams. Note that where ScratchMaths was not taught by the person teaching mathematics there may have been a barrier to the teacher mediation mechanism posited in the ScratchMaths theory of change.

Table 24 below shows when ScratchMaths was taught, as reported by survey respondents. At the start of the trial the ScratchMaths team collected email addresses from all teachers who were subsequently emailed and invited to participate in the survey. Where individual emails were not available, general school email addresses were used. As detailed in Table 5 surveys were conducted in the Summer terms of 2016 and 2017.

Seventy two per cent of responding teachers reported that ScratchMaths was taught outside mathematics lessons in Y5 and 86% reported this to be the case in Y6. Further, 24% in Y5 reported that ScratchMaths took place in time that was additional to regular computing lessons and 54% reported this in Y6. Thus, it appears in many schools additional curriculum time was allocated to ScratchMaths beyond that normally allocated for mathematics and computing. So, arguably, more time for computing and mathematical learning took place in the Wave 1 schools than was usual and presumably more than in the Wave 2 schools.

Table 24: Timing of ScratchMaths lessons according to teachers responding to the survey

	Frequency		Percentage	
	Y5	Y6	Y5	Y6
Taught in timetabled computing lessons	26	11	58	31
Taught in mathematics lessons	8	4	18	11
Taught in additional ScratchMaths lessons	11	19	24	54
No response	0	1	0	3
Total	45	35	100	99 ³⁰

Use of materials

Both the interview and survey respondents reported using all or most of the core materials. In interviews, all Y5 teachers reported using most or all extension materials. However, only two teachers reported doing so in Y6. Reasons for this, as far as they are known, appear to be similar ones to those generally affecting implementation (see 'Supporting conditions and barriers below'). Presentations were used by all interviewees.

There was a general pattern of the amount of material used declining over the course of each year and from Y5 to Y6. In Y5, 91% of 43 respondents to the question had used module 1 materials and the same for module 2, with 86% using module 3. However, whereas for module 1 and 2, the minimum percentage that reported using any of the four investigations was over 80%, for module 3, 50% reported using investigation 3 and 43% investigation 4. In Y6, of 31 respondents to the question, 84% reported using module 4 (the first module taught in Y6), 61% module 5 and 45% module 6. The amount of time teachers spent using materials is considered further in the discussion of fidelity below.

Where variation in use of materials was reported in interviews, this was generally described as minor, for example 'tweaking' a presentation, or using all but one activity in a module.

Time spent teaching ScratchMaths

The table below provides survey data on the amount of time spent teaching ScratchMaths for the respondent sample over the school year. It shows that the amount of time spent teaching ScratchMaths was proportionally less in Y6 than in Y5.

Table 25: Time spent teaching ScratchMaths

	Y5 Frequency	Percentage	Y6 frequency	Percentage
n=	36		31	
20 hours or more	18	50	5	16
12 to less than 20 hours	16	44	18	58
Less than 12 hours	2	6	4	13
Missing response	0	0	4	13

³⁰ Rounding error

Supporting conditions and barriers to implementation

Interviewees were asked about supporting conditions and enabling factors and the seven factors named by the nine respondents were given as:

- Senior leader support.
- Computing a priority in the school.
- Positive culture for computing.
- Access to good/suitable equipment.
- Quality of the materials.
- Quality of the professional development.
- Taking time to work through materials before using them.

The importance of professional development is also suggested by analysis of the relationship between degree of attendance at PD events and use of materials in Wave 1 Y5 (positive correlation between number of days' attendance and hours teaching of ScratchMaths to Y5 $r=0.31$, $p<0.05$); however, a similar relationship was not found in Wave 1 Y6 (no significant correlation between number of days' attendance and hours teaching ScratchMaths to Y6 $r=0.23$, $P>0.05$). Interpretation of the difference between the correlation found for Y5 and Y6 is confounded by the lower PD attendance in Y6.

In the Wave 1 Y5 and Y6 survey, teachers were asked about the barriers to implementing ScratchMaths. The list of possible responses used in the survey item was generated from Y5 interviews and ScratchMaths team evaluation and feedback. In Y6, additional responses were added based on feedback from participants and local hub leads to the ScratchMaths team. Responses are reported below. Participants could select any that applied to their context and so the percentage shown is of survey respondents in each case.

The issue of mathematical demand for learners and, more generally, the challenge of materials was cited by a similar percentage of interview respondents overall. The issue of using materials with pupils with a range of prior knowledge of Scratch and with prior mathematical attainment was a concern for a significant minority. In addition, from interviews suitability for low-attaining pupils was also a concern. From interviews, technical difficulties appeared mainly to be around using Scratch online.

It is notable that 69% of Y6 teachers reported KS2 maths standard assessment tests (SATs) as a significant pressure, as well as timetabling pressures (57%). These findings echo the reasons given by schools for not continuing implementation discussed above. For 20% of Y6 respondents, lack of knowledge of the Y5 ScratchMaths curriculum was an issue; this issue arises because only a minority of the Y6 teachers had attended Y5 PD then implemented ScratchMaths in Y5 before attending Y6 PD. An arguably better arrangement was adopted in some schools of the same teachers attending PD focused on Y5 materials, teaching ScratchMaths to Y5, and then attending the PD focused on Y6 materials and then teaching ScratchMaths to Y6. However, this was only likely to be possible in schools that, as policy, had teachers carry through from Y5 to Y6 or where computing (or specially designated ScratchMaths lessons) were taught by a specific teacher.

Table 26: Barriers to implementation ordered by mean percentage across Y5 and Y6 where applicable

	Frequency Y5	Percentage Y5	Frequency Y6	Percentage Y6	Mean percentage Y5 & Y6
Pupil difficulties with mathematics	22	47.8	8	22.9	35.3
Technical difficulties	20	43.5	9	25.7	34.6
Differentiating materials	14	30.4	8	22.9	26.7
Own lack of confidence and/or knowledge of programming	5	10.9	10	28.6	19.8
Access to technology	10	21.7	4	11.4	16.6
Other	6	13	1	2.9	14.5
Lack of sufficient time for lesson preparation	6	13	5	14.3	13.7
Pupils having previous experience with Scratch	8	17.4	3	8.6	13
Pressure to prepare pupils for KS2 maths tests (Y6 ONLY)	N/A	N/A	24	68.6	N/A
Lack of knowledge of Y5 ScratchMaths curriculum (Y6 ONLY)	N/A	N/A	7	20	N/A
Timetabling pressures (Y6 ONLY)	N/A	N/A	20	57.1	N/A

Caution is advised about interpretation of differences between Y5 and Y6, because of the small sample size, which is not composed of identical respondents (or from the same schools) in the two years. However, the percentage reporting 'pupil difficulties with mathematics' in Y5 is over twice that in Y6. This may be an indication that the perceived mathematics demand for Y5 pupils was greater than for Y6. The reduction in 'technical difficulties' is likely due to the original suggestion of using the on-line version of Scratch. Due to connectivity issues, interviewees reported greater success when using Scratch off line. There is an increase in the percentage reporting lack of confidence and/or knowledge of programming from Y5 to Y6. This may be an indication that teachers who first engaged with Scratch in Y6, were using materials that assumed a level of prior knowledge. With regard to 'pupils' having previous experience with Scratch, the ScratchMaths team conjectured that if pupils already knew how to programme this might lead to less engagement in the materials. Participants were invited to provide comments for 'other' responses, and analysis indicates in all but one case that these related to the other categories in the table. The one exception was a statement by one respondent that ScratchMaths could not be taught if the respondent was not at the school. Although not stated in responses to this question, other comments suggest that they considered their relative level of skill in computing as much higher than other teachers in the school, and essential to being able to use ScratchMaths materials.

It is notable that 68.6% of respondents in Y6 identified SATs pressures as a barrier to implementation. This is also confirmed by responses to the ScratchMaths team regarding reasons for not implementing or attending in Y6 by those schools that did not do so and provided information. In addition, timetabling pressures also appeared (from interview responses) to have an element related to SATs pressure.

Teacher views of the intervention

Respondents in the SHU survey, the ScratchMaths team post PD survey (see Appendix L) and in telephone interviews were, in general, positive about the quality of professional development. Those new to coding in particular reported that the PD was appropriate for them and in general the amount of time for PD events was appropriate.

Similar views were expressed about the materials and these were compared favourably to alternatives. An example of the views of teachers who were favourable about ScratchMaths is given below.

The resources are brilliant for teaching the children from basics the steps involved and why you use them. We do have other resources we've used for coding and though the children can complete the exercises, they often don't understand the underlying principles. With ScratchMaths resources, they do. (Wave 1 Y6 teacher, Interview)

A contrasting view which represents both the positive attitude of many respondents to the materials and programme but also the challenges in implementing them is provided below.

The materials/resources are excellent and it is a well thought out and designed programme. Its limited use to us at the moment is more a reflection of where we are at the moment as a school in terms of our computing ability. The many problems and difficulties we encountered are not really a criticism of ScratchMaths but in our experience this year we found it to be extremely demanding on time teachers spent preparing lessons (going through projects to ensure their own subject knowledge was up to scratch - no pun intended - creating individual files with children's names (I experimented with a few offline/online versions and this was not ideal but the least troublesome), the programme got way too difficult for all but one of my pupils and most sessions regrettably turned into more a case of me giving instructions and children following rather than children learning and discovering for themselves and the sessions were dramatically more computing focused as opposed to mathematics focused. (Y6 survey respondent)

For some teachers interviewed, learning Scratch themselves was challenging, and it may be that this is an explanation for lower fidelity in Y6 where time pressures were more acute and schools less willing to give teachers the time needed to work through materials themselves.

Whilst materials were well-regarded by most teachers surveyed or interviewed, a theme for some was that lower-attaining pupils found it challenging to access all the materials. Table 27 presents responses to the survey question as to whether all pupils found materials accessible. It is likely that some of the schools who did not continue from Y5 to Y6 also found that their pupils found it difficult to access materials.

Table 27: Y5 and Y6 teachers' perceptions of pupils' ability to access materials

	Wave 1 Y5		Wave 1 Y6	
	Frequency	Percentage	Frequency	Percentage
Yes	30	65.2	22	63.3
No	14	30.4	8	22.9
Left Blank	2	4.3	5	14.3

As can be seen from the table, a similar proportion of interview respondents in both years highlighted issues for lower-attaining pupils (30.4% in Y5, and 22.9% in Y6).

Table 28 below summarises teachers' responses to the survey as to whether they would use the materials again if they were teaching the following year

Table 28: Y5 and Y6 teachers' future use of materials

	Y5		Y6	
	Frequency	Percentage	Frequency	Percentage
All or most of the materials	27	61	9	31
Some of the materials	15	34	17	58
No or few materials	2	5	3	10

The responses about future use of materials confirm that in general teachers had a more positive view or experience in Y5 than Y6.

Suggestions about improving materials tended to be specific to individuals, although two areas were highlighted by a number of interviewees: firstly, the suggested amount of time for use of materials should be longer as this was more realistic, and secondly the overall level of challenge should be reduced.

Teachers were asked about their overall views of the project. As can be seen from the table below, responses were overwhelmingly positive in Y5 although less so in Y6.

Table 29: Overall view of the ScratchMaths project

	Wave 1 Y5		Wave 1 Y6	
	Frequency	Percentage	Frequency	Percentage
Very positive	20	43.5	3	8.6
Positive	18	39.1	18	51.4
Neither positive nor negative	2	4.3	9	25.7
Somewhat negative	2	4.3	0	0
Very negative	1	2.2	0	0
Left Blank	3	6.5	5	14.3

As reported above, a minority of teachers considered the material to be overly challenging for some pupils. However, this was not universal, as the following quote indicates:

We've seen the pupils that were in the lower group for maths show a real talent and enthusiasm for coding. (Wave 1 Y5 Teacher, interviewee)

In general, teachers stated that pupils were positive about ScratchMaths materials, although tending to view it as computing and not mathematics. Many pupils were perceived to be enthusiastic, as exemplified by the following report from one teacher:

Seeing children have a lesson where they want to carry on after the lesson - "can I stay in at lunchtime and do more of these". (Wave 1 Y6 teacher, interviewee)

Teacher mediation

During the second year of the intervention, the ScratchMaths team, in reviewing the theory of change, highlighted the importance of teacher mediation. This was addressed, following discussion with the ScratchMaths team, in the Y6 interviews, and in the Y6 survey, by asking whether ideas from

ScratchMaths informed mathematics teaching outside of ScratchMaths lessons. Survey data are presented below, in Table 30, concerning the question as to whether ideas from Scratch or ScratchMaths had been used in mathematics lessons other than ScratchMaths ones. Of those who responded to the question, just over half had rarely or never used ideas from Scratch or ScratchMaths in other mathematics lessons, indicating that one possible marker of mediation was not frequently enacted for half the teachers. Given the limits of the survey, further data on how and in what way teachers used ideas from ScratchMaths in other lessons was not collected, but could be a subject for further research. The implications of most teachers not using ScratchMaths ideas in other lessons for the theory of change are discussed below in the conclusion where findings are interpreted.

Table 30: Reported influence of Scratch and ScratchMaths on mathematics teaching

Wave 1 Y6			
	Frequency	Percentage of those responding to the survey	Percentage of those responding to the item
Often	1	2.9	3.6
Sometimes	12	34.3	42.9
Rarely	8	22.9	28.6
Never	7	20.0	25.0
No response to the item	7	20.0	N/A

Fidelity

In the section on implementation and process evaluation methods, fidelity criteria were presented, as was the approach to assessing fidelity. Further details are provided in Appendix K. These criteria are applied to the implementation data by considering each dimension in turn and then the overall fidelity profile. Of the 55 intervention schools, teacher survey data was collected from 36 schools in 2015/16 (Y5) and 31 schools in 2016/17 (Y6). We obtained complete Y5 and Y6 survey data for 27 of the 55 intervention schools.

Professional development attendance fidelity

Two sets of data were collected for ScratchMaths PD attendance: from IoE registers and from the two IPE surveys. The IoE data for this dimension of fidelity are preferred as more reliable due to being collected at the time of PD attendance. Details of the differences are found in Appendix J, and Appendix K has details of the analysis of fidelity for the subsample of schools where there are fidelity data across all five dimensions for both Y5 and Y6.

Drawing on data presented above, on professional development implementation, the following teacher-level fidelity patterns are found (see Table 31). This is described as 'core fidelity' as it was possible for a teacher to meet the fidelity criteria if there was a process internally for school-level PD. This could only be determined through inspection of survey data.

Note that the 'low fidelity' data is an estimate based on assuming a potential attendance of two teachers per school as in the project design. Calculating accurate attendance fidelity is challenged by withdrawal, variations in size of schools, rotation of teachers attending and staff changes. There is an increase in low fidelity from 19 teachers in Y5 to 57 teachers in Y6, and this can be explained largely through school withdrawal.

Table 31: Teacher-level core fidelity considering ScratchMaths attendance data

	Y5	Percent	Y6	Percent
n=	110		110	
High	63	57%	41	37%
Medium	28	25%	12	11%
Low	19	17%	57	52%

Source: ScratchMaths attendance data

The pattern found in the full sample, using ScratchMaths team attendance data, is lower attendance fidelity in Y6. However, this is not reflected in the survey responses as shown in Table 32. Thus, there is sample bias in the survey responses and this is particularly marked in Y6.

Table 32: School fidelity using ScratchMaths attendance data for survey respondents only

	Y5	Percentage	Y6	Percentage
n=	36		31	
High	24	67	27	87
Medium	9	25	0	0
Low	3	8	4	13

Source: ScratchMaths attendance data

Technology fidelity

In Y5, 35 out of 36 schools met the high-fidelity criterion of a ratio of two pupils per computer and in Y6, 30 out of 31 schools (for the subsample of 27 for which there was data for both years, all schools met the criterion). High fidelity here is not surprising as access to technology was an aspect of eligibility criteria.

Module use fidelity

The module use fidelity dimension had two components: fidelity in Y5 and fidelity across Y5 and Y6.

For Y5, 72% of the 36 schools' composite survey responses reported teaching sufficient material from all three modules (high fidelity), 25% had taught sufficient material from two modules (medium fidelity), and 3% had taught material from fewer modules than this (low fidelity).

While there was no criterion for Y6 alone, comparative data from the Y6 survey show that 42% report teaching from all three modules, 23% from two modules and 35% from fewer than this.

For Y5 to Y6 progression, it is only possible to consider schools where there were responses for both Y5 and Y6 (n=27) in the on-treatment sample. For this sample, fidelity is shown in the table below.

Table 33: Module use fidelity for restricted sample

Fidelity	Number of modules	Frequency	Percentage
High	5 or 6 modules	17	63%
Medium	4 modules	2	7%
Low	3 or less modules	8	30%

However, if one assumes that those who were highly engaged were more likely to complete the survey in both years, and there is evidence of this from attendance patterns for survey respondents, then it suggests that the overall fidelity of the original sample could be as low as 31%. Further, this may be an over-estimate given the data reported in the section on implementation module use that some survey respondents reported that there was only one class in their school following ScratchMaths in Y6, and this is not explained by responses from one-form entry schools.

Time spent teaching ScratchMaths fidelity

Table 25 above, in the section on implementation, presented the amount of time spent teaching ScratchMaths. Because the fidelity criteria are concerned with combinations of time spent over two years, fidelity can only be reported for the restricted sample of 27 schools.

For Y5 and Y6, the two years need to be inter-related. This is shown in Table 34 below.

Table 34: Y6 and Y5 time spent on ScratchMaths matrix

Y5	Y6			
	Less than 12	12 - 20	20+	Missing
Less than 12	0	1	0	0
12-20	1	6	3	2
20+	2	8	2	2

This identifies 10 schools with high fidelity (20+ hours in Y5 and 12+ in Y5) and nine schools with medium fidelity.

Progression fidelity

The progression fidelity criterion was formulated as a binary: whether the intended progression of modules was followed. The large majority of respondents reported this affirmatively, as Table 35 shows.

Table 35: Y5 and Y6 module progression fidelity

	n	Yes	No	Missing
Y5	36	35 (97%)	1 (3%)	0
Y6	31	27 (87%)	3 (10%)	1 (3%)

Overall fidelity

Table 36 below summarises, for the subsample of 27 schools, overall fidelity combining all five fidelity dimensions for which there are data from attendance records and both the Y5 and Y6 surveys. Note it is not possible to provide an overall measure of fidelity for other schools who responded to only one survey, due to missing data.

Table 36: Summary fidelity analysis across five dimensions for all schools with complete survey data

	Y5 (n=27)	Y6 (n=27)	Y5 and Y6 (n=27)
High Fidelity	7 (26%)	5 (19%)	5 (19%)
High/Medium	24 (89%)	13 (48%)	13 (48%)
Low	3 (11%)	14 (52%)	14 (52%)

Variation in implementation and fidelity

It was reported above that, at most, only 71% of schools could be considered to have sustained participation over the two years and some of these had low fidelity. School-level variables were compared for schools that were classified as having sustained participation and those that were not. Significance tests on comparison of means identified no significant differences in relation to KS2 maths level 4+ attainment, or numbers of FSM, English as Additional Language (EAL) or Special Educational Needs (SEN) pupils between schools with sustained participation and those without, in the Wave 1 schools. This means that variation in implementation does not appear to be due to observed school variables.

Further, the relationship between attainment at school level and FSM (as a proxy for demographic differences) and hours spent teaching ScratchMaths and attendance was examined. No correlations were found. Thus, there are no discernible differences in implementation by school characteristics.

The ScratchMaths team asked staff in schools who had stated they were not continuing with the programme for the reasons for their decision. Those who responded reported: staffing pressures (changes and illnesses); prioritising KS2 tests; general pressure on schools; and computing not being a priority or being deprioritised in the school.

In addition to schools not continuing to participate, as reported above, data from surveys indicate that in both Y5 and Y6, fewer than half the schools were likely to have implemented ScratchMaths across all fidelity criteria at high or medium level.

Fidelity analysis identifies a pattern of lower fidelity in Y6 than in Y5 in relation to attendance, module use and time spent teaching ScratchMaths. With regard to module use, this was in spite of a revision of the criteria to anticipate that schools might only teach two out of three modules in Year 6. The most important reason for this was the systemic pressure of KS2 assessments in many schools, as well as pressures on staffing.

The evidence is more mixed as to whether intrinsic aspects of the intervention itself also led to low implementation in some schools. In general, schools who continued to participate were positive, or very positive, about the professional development and materials. However, there are indications that may suggest barriers to implementation in the current design of ScratchMaths.

Teachers with little knowledge of coding reported needing to spend a lot of time working through materials themselves. Where they were able to spend this time, this was reported as leading to good professional development outcomes. However, if time for additional planning was not available, using the material was more challenging. For those Y6 teachers who had not taught ScratchMaths in Y5 (40 of 65 who attended at least one Y6 PD event), there was the additional challenge of teaching material that progressed from concepts and skills already taught to pupils. Arguably, the two days' summer professional development were not sufficient for these teachers both to 'catch up' on Y5 ScratchMaths concepts and become familiar with Y6 material as well as develop their own programming skills.

Outcomes

Professional development outcomes

Teachers reported improved computing skills and computing teaching knowledge. For most teachers prior to engagement in ScratchMaths, their knowledge of Scratch programming was limited, with a number of interviewees stating that developing their programming skills and knowledge of Scratch was the most positive aspect of the project, for example:

[SM has had a] Big impact on the coding skills of both [the] children and me (Wave 1 Y5 teacher interviewee)

Improved skills and knowledge were linked to increased confidence to teach ScratchMaths. A minority of respondents also reported changes in beliefs about teaching mathematics and/or computing.

In addition, details of favourable responses to training in post PD evaluation feedback related to professional development outcomes are found in Appendix L.

Teacher perceptions of pupil outcomes

In the interviews, teachers were asked about the effect of using ScratchMaths materials on pupils' computing skills, computational thinking, problem-solving and mathematics. In relation to the first three aspects, 16 out of 18 teachers interviewed stated that they believed ScratchMaths had a positive effect.

In relation to outcomes on mathematics, responses were more mixed. In Y5, a minority of teachers were positive about outcomes on mathematics but highlighted that this was in relation to specific content. In relation to Y6, similarly, less perceived impact was reported than with regard to computing, computational thinking and problem-solving. However, more examples were offered of changes to mathematical learning than in Y5. This accords with the relative difference in focus in mathematics in the Y5 and Y6 materials.

A number of teachers also commented on improved resilience of pupils:

The children are becoming more resilient and more logical in their thinking. (Wave 1, Y5 teacher, interviewee)

Formative findings

A number of interviewees reported that the half-day PD sessions were very helpful and others that a better PD model would be for professional development to be spread over the year rather than two full days in the Summer term before starting to use the materials.

Some teachers with little experience of Scratch found learning to code difficult themselves. Some had the confidence to adopt a 'learning together' approach with their pupils and/or were able to commit additional time to work through materials. However, it may be that some initial PD to look at Scratch basics more gradually would be useful for some teachers.

The level of challenge of the materials was too high in the view of some teachers and furthermore, given the amount of material per year, it was difficult to use it all in the time available. Related to this was the issue that where materials were used effectively and as intended, teachers needed to spend a lot of time planning. This was particularly true for teachers with less prior experience of Scratch. One teacher, who was positive about ScratchMaths and intended to use the materials in the future, stated that they would spread the material across Y4 to Y6 (three years not two) as this would be more achievable.

The programme was in general experienced as a computing project rather than a computing and mathematics project by teachers and, from their reports, by pupils too.

Teacher mediation between ScratchMaths learning and mathematics appears limited, even at the simple level of using Scratch or ScratchMaths ideas in other mathematics lessons. This suggests that greater attention needs to be paid to this aspect of the programme in professional development, and this may require additional PD time.

It is notable that 24% of Y5 survey respondents and 54% of Y6 respondents reported that ScratchMaths was taught in additional ScratchMaths lessons. Where schools were able to organise this and were willing to devote additional curriculum time to ScratchMaths, fidelity was higher. Potentially, this could be suggested as a good way of implementing ScratchMaths, at least to begin with while teachers were becoming familiar with the materials.

In general, schools found it easier to use Scratch offline than online.

There are considerable challenges to undertaking a computing professional and curriculum development programme in Y6 in the current accountability context experienced in English primary schools in which KS2 maths and English tests are the main focus in Y6, and in some schools an overwhelming focus.

Control group activity

The control group consisted of a Wave 2 cohort of pupils and their teachers who were in Y5 in 2015/16 and Y6 in 2016/17. Additionally, in the control schools the 2016/17 Y5 cohort experienced ScratchMaths.

Meaningful data are only available from the Wave 2 Y5 teachers in 2015/16 and 2016/17 and not the Y6 2016/17 teachers - see Appendix J.

In the impact analysis section above, the balance across pupil participants was presented in relation to a range of variables. Data on prior knowledge or use of Scratch was collected firstly in a pre-PD survey by the ScratchMaths team (and as stated in the IPE methods section then provided to the evaluation team) and, secondly, through the Y5 and Y6 teachers intervention and control surveys. These data indicate that the control schools and teachers were not markedly different from intervention schools in relation to prior knowledge and use of Scratch. In the IPE surveys, teachers were also asked about other activities that might have influenced outcomes. Again no relevant differences were found between intervention and control schools

In addition, data from the survey and interviews with Wave 1 Y6 teachers indicate that due to concerns with SATs, teachers found it hard to use ScratchMaths material even though they had committed to doing so. Thus, it seems unlikely that Y6 teachers in control schools would use additional material intended for Y5 when not required to do so and when requested not to use the material.

Conclusions

Key conclusions

1. There is no evidence that ScratchMaths had an impact on pupils' KS2 maths outcomes. This result has a very high security rating.
2. Children in ScratchMaths schools made additional progress in computational thinking scores at the end of Year 5, compared to children in the other schools. The additional progress was higher for children who have ever been eligible for free school meals.
3. Many schools did not fully implement ScratchMaths, particularly in Year 6. High fidelity to the intervention was found in 44% of schools in Y5 and 24% in Y6. Implementation was enhanced where schools provided teachers with time to work through materials.
4. Teachers viewed ScratchMaths as a good way of addressing aspects of the primary computing curriculum, good for improving Scratch programming skills, good professional development, and good for its high quality materials. Five teachers voiced concerns that the lower-attaining pupils needed additional support or adaptation of materials to fully access all ScratchMaths content.
5. Participation in professional development and the use of materials is potentially a very low-cost per pupil option to enhance non-specialists' knowledge and skills to teach aspects of the primary computing curriculum in a manner that is suitable for boys and girls.

Interpretation

Here, each of the evaluation research questions are considered in turn, as are issues of scalability, and the findings are summarised and discussed. For concision, definition of terms and detail of the trial and, for example, the CT tests are not given here but can be found elsewhere in the report.

RQ1: What has been the effect of the intervention on the development of pupils' mathematical skills as measured by a randomised control trial?

The intention-to-treat analysis found no evidence that the ScratchMaths intervention had an impact on the development of pupils' mathematical skills as measured by KS2 maths tests. This was found when KS1 maths attainment was controlled for and when attainment in both the KS1 maths and CT tests were controlled for. No evidence of impact was found when the analyses were restricted to a subsample of schools identified as having high or medium/high fidelity to the ScratchMaths intervention, nor when individual KS2 maths test papers were considered separately. It is possible that impact on mathematics attainment might be delayed and this could be assessed through analyses of KS4 maths attainment when this pupil cohort reaches this stage in Summer 2022.

Given there was a positive effect on computational thinking, but not on mathematical attainment, interpretation of this finding in relation to the intervention theory of change is discussed in relationship to RQ5, below, focused on the relationship between computational thinking and mathematical attainment.

RQ2: How can computational thinking be measured?

A short online test with binary scoring was developed based on Beaver/Bebras items (an international computing competition) The use of these types of items was suggested by the ScratchMaths team, although, they later expressed reservations about the choice of items included in the test for use with Y5 pupils (as stated earlier, the ScratchMaths team were provided with a copy of the test after it has been used in the main trial). Their reservations about the choice of items related to changes to the computing national curriculum in England in 2014 and to the content of ScratchMaths. They suggested that items should have been selected from Beaver/Bebras tests for older children than those involved

in the project. However, analysis of test outcomes suggests it is of an appropriate level of difficulty for this age of children. This is not surprising given it was modelled on Bebras challenges aimed at KS2 primary age children. The test was successful in detecting a change in computational thinking and has appropriate psychometric properties, as well as being efficient to administer. The test is appropriate for use in similar trials and evaluations, and Beaver/Bebras items could be used in tests designed for children of other ages.

RQ3: What correlation exists between measured computational thinking and mathematics attainment?

Table 37 below summarises correlations between scores on the computational thinking test and mathematics attainment. For KS2 maths, the correlation with overall (raw) mathematics score is reported. Correlation coefficients with sub-scores for Paper 1 (arithmetic), Papers 1 and 2 (reasoning 1 and 2) are +0.39, +0.46 and +0.45 respectively. Correlations for subsamples of intervention and control groups are reported in Table 16 above in the Impact Evaluation section. The one observed difference was that the correlation between the CT test and KS2 maths scores was stronger for the control group ($r=+0.50$) compared with the ScratchMaths intervention group ($r=+0.41$).

Table 37: Summary of Pearson's correlation coefficients found between CT test scores and KS1 and KS2 mathematics scores

Sample	Key Stage	CT test and Mathematics correlation
Pilot (231 Y6 pupils, from 6 schools)	KS1 (Y2)	+0.36
	KS2 (Y6)	+0.45
Trial (intervention and control) Y5 CT	KS1 (Y2)	+0.49
	KS2 (Y6)	+0.47

Thus, there is a medium to strong correlation of CT test scores with KS2 mathematics scores in both the pilot and main samples. In addition, the correlations with KS1 grades are statistically significant but for the pilot data are slightly attenuated compared with the KS2 relationships. The medium to strong correlations observed with mathematics attainment suggest that there are shared core components being measured by KS2 mathematics and CT scores.

RQ4: What has been the impact of the intervention on the development of pupils' computational thinking?

The ScratchMaths intervention led to a statistically significant positive impact on pupils' computational thinking in Y5 with an effect size of +0.15 standard deviations (sds) and an estimated statistical power of 60%. Impact was greater for FSM pupils where an effect size of +0.25 sds was observed.

As noted earlier, a recent evaluation of code clubs (Straw, Bamford and Styles, 2017) found no change in computational thinking as measured by a similar test (with some identical items). Code clubs involve pupils spending additional time learning Scratch, Python and HTML programming. As noted in Appendix F, the way the code club measure was scored was different and so it is possible that the difference is due to the more complex assessment method. However, the contrasting positive result for ScratchMaths indicates that the effect is a result of the intervention rather than learning to code in general.

RQ5: What conclusions can be drawn about the relationship between mathematical thinking and computational thinking from the quantitative analysis?

The observed correlations between mathematics and computational thinking support the contention that these are related. However, the contention that computational thinking is a subset of mathematical thinking, as suggested by some (for example, Wing, 2008, 2011), is not borne out. Differences in computational thinking between control and intervention pupils were not matched by differences in mathematics attainment, which suggests some or a combination of the following most likely possible interpretations:

1. Although the CT test and mathematics attainment draw upon some core mathematical competencies, the Y5 ScratchMaths curriculum impacted upon competencies unique to computational thinking or at least not tested in KS2 mathematics assessment; this suggests the need to revise and make more precise the programme theory of change (Figure 1) to specify which aspects of computational thinking need to be increased in order to impact on mathematical attainment.
2. The theory of change is tenable but changes in computational thinking were not sustained in Y6. and so did not impact on Y6 KS2 mathematics test attainment. In general, schools that administered the CT test did continue to attend ScratchMaths and use the Y6 materials). However, given that pupils continued to follow the intervention in the high/medium fidelity schools and the on treatment analysis did not identify a positive effect, this proposition appears unlikely. Nevertheless it is possible that the effect on computational thinking was due to particular materials or activities in the Y5 ScratchMaths programme for reasons that are not apparent.
3. The underlying programme theory of change is flawed and changes in computational thinking do not lead to changes in mathematical attainment in the way posited.
4. There was not sufficient teacher mediation for changes in computational thinking to lead to changes in mathematical attainment, given there is evidence from the process evaluation that teacher mediation was limited.
5. Changes in computational thinking were sustained but were not of a sufficient size to lead to changes in mathematical attainment.
6. Changes in computational thinking were sustained and did lead to changes in mathematical attainment but these changes were minor compared with the effect of other mathematics teaching during Y6, for example KS2 SATs preparation. Or, similarly, the positive changes were balanced by adverse effects of engagement on KS2 mathematics attainment, due to teacher and pupil attention to and engagement with computing rather than mathematics (adverse here is meant in a narrow sense related to KS mathematics test outcomes only).

A further possibility, and related to point 1, is that the KS2 mathematics test, even though two of the three papers are designated as 'problem solving', does not assess well the type of mathematical thinking that is closest to computational thinking. However, given the importance of KS2 mathematics attainment to schools, then this suggests the programme needs developing to connect change in computational thinking to the type of mathematical thinking tested in KS2 tests. This suggests that greater attention needs to be given to teacher mediation.

The ScratchMaths team also pointed to differences between the 2017 KS2 mathematics test content and the 2016 test, the latter test had informed their design. They suggested that the 2016 tests might have been a more appropriate test for the content of ScratchMaths. The overall time spent engaging in ScratchMaths involves Scratch programming skills, computational thinking and computing knowledge as well as the mathematical knowledge in the contexts used for ScratchMaths. Thus, the amount of curriculum time that involves engagement in mathematics directly is small relative to usual mathematics curriculum time. So, for engaging in computing to impact on mathematics to any great extent or for a case to be made for mathematics to be practised in the context of computing rather than other

curriculum subjects, then the posited relationship between computational thinking and mathematical thinking is important.

It is also noteworthy, as reported above, that in a majority of schools ScratchMaths was taught in either computing lessons or in additional ScratchMaths lessons. Thus, arguably there was greater curriculum time for mathematics or mathematics-related activity in the intervention schools than in the control schools. In particular, 24% of teachers responding to the Wave 1 Y5 survey and 35% responding to the Wave 2 Y6 survey report ScratchMaths being taught in additional lessons. If other schools were to allocate additional time to mathematics in Y5 and Y6 then alternatives to ScratchMaths should be considered. However, if the priority is to increase time for computing then ScratchMaths may be a good option given teachers' positive view of outcomes in relation to pupil programming and its measured effect on computational thinking.

RQ6: To what extent does the design and delivery of curriculum materials and professional development and the associated materials fit with the current knowledge base on effective professional development in relation to mathematics teaching/computing?

The intervention had a number of features of effective professional development: connecting work-based learning and external expertise; potentially rich professional learning opportunities; collaborative learning; the creation of professional learning communities between schools; and a clear focus. This aligns with the current knowledge base on effective professional development (for example, Stoll, Harris and Handscomb, 2012; Cordingley et al, 2015). Similar features have also been specifically identified as important in professional development for mathematics teachers in England (ACME, 2016; Back et al., 2009). The professional development and introduction to materials also accord with principles recommended for high-quality professional development for computer teachers (for example, Naace, n.d., given Naace's involvement in the study then this is, perhaps, to be expected.). Given the time constraints, the approach to professional development was efficient and effective for teachers with varied levels of prior experience of Scratch. However, for those teachers who were unfamiliar with Scratch the amount of professional development was possibly too short.

RQ7: What are the teachers' views on the effectiveness of the professional development?

The teachers, in general, viewed the professional development as effective. Evidence for this was consistent across ScratchMaths team post-PD event surveys, teacher surveys and interviews, and also consistent for both Y5 and Y6 professional development. The expertise of the ScratchMaths team was appreciated. However, given that participation in PD was lower in Y6, and given the issues of survey sample bias, it is likely that the survey data do not represent fully the views of those teachers who found a considerable level of challenge in relation to their starting points in relation to Scratch programming and teaching Scratch.

RQ8: Were there any barriers to implementing ScratchMaths, or were there particular conditions that needed to be in place for it to succeed?

Implementation of ScratchMaths was enhanced by schools' release of teachers and provision of time to work through materials, particularly for teachers who were less familiar or unfamiliar with Scratch, and schools' willingness to support computing teaching in Y6 given pressure of KS2 test requirements. Barriers to implementation were systemic issues related to KS2 accountability pressures, staffing issues including confidence and changes, and, reported by some teachers, the level of challenge of materials for some pupils.

RQ9: In what ways can the professional development delivery and materials be improved?

Possible ways to improve professional development delivery are:

- Spread the professional development events more evenly over the year.

- Address issues of teacher mediation between ScratchMaths and mathematics more explicitly in the professional development; this may require an increase in professional development time.
- Consider the needs of teachers who found learning Scratch particularly challenging, potentially with some preliminary training sessions.

Possible ways to improve materials:

- Consider the challenge of materials for low attaining pupils.
- Reduce the amount of materials per year, possibly by spreading the programme over three years.
- Design materials for use offline rather than online.

The programme was in general experienced as a computing project rather than a computing and mathematics project, so the mathematical content of the programme needs further developing or otherwise made more transparent to teachers and learners.

The overall approach to professional development and organisation of materials as well as the pedagogical approach has potential to inform professional and curriculum development focused on computing and programming generally.

Scalability

In addition to the research questions, a further objective of the evaluation was to investigate issues of scalability. Barriers to implementation were summarised above. Given that many of these are structural barriers and difficult to overcome through intervention design, some caution is needed about scaling the intervention.

Further, it is notable that overall attendance per teacher and school was lower for Wave 1 Y6 than Y5 (with SATs pressure cited as a reason). In addition, attendance and material use were lower again for Wave 2 Y5 (who experienced ScratchMaths as per the waitlist design). PD for these teachers was led by ScratchMaths local coordinators and it may be that this was less attractive for teachers and schools than PD led by the ScratchMaths team. This may be a challenge to overcome if the programme is extended or if an effectiveness trial is undertaken with delivery by local providers.

However, given the overall process evaluation findings and the view of the ScratchMaths delivery team, it appears an important issue in the reduction of fidelity between Y5 and Y6, and so for the programme overall, is the pressure in Y6 to focus on national tests (see table 20 and discussion). Given this, a revision of ScratchMaths might be easier to implement if it focused on Y4 to Y5 (for example, by splitting the current Y5 materials over these two years) or re-designing the Y6 materials as a Y7 or Y8 project, or alternatively as a Y6-Y7 transition project to be studied after the KS2 tests.

Value of ScratchMaths to schools

In its current form, ScratchMaths cannot be advised as a means of raising mathematical attainment. However, for schools where there is a perceived need to develop teachers' computing skills and there is support from senior leaders to give time for planning and release, then ScratchMaths is potentially a very low-cost (per pupil) and effective form of professional development.

Further, ScratchMaths achieved a measurable increase in computational thinking whereas this has not been found in the Code Clubs evaluation (Straw, Bamford and Styles, 2017). This increase was more pronounced for FSM pupils. Although outcomes in computing are not a central focus in English primary school accountability measures, increasingly computing and computational thinking are viewed as an important aspect of education (for example, included as an English Baccalaureate subject at the end of KS4). ScratchMaths can be advised as an effective way of developing computational thinking.

Moreover, teachers considered ScratchMaths to be an effective way of learning to programme, although this was not measured in the trial.

Limitations

Potential spill-over

The waitlist design meant that Y5 pupils in the control schools received the intervention during the second year of the trial (2016/17). There is, then, a risk of potential spill-over from those Y5 teachers and classes to the control Y6 teachers and classes. Data investigating the possible spill-over were collected by survey of teachers as part of the implementation and process evaluation, and there are no indications that spill-over occurred, although there are missing data for Y6. However, as noted above, given challenges for the intervention schools in implementing the programme in Y6, it seems unlikely that Wave 2 Y6 teachers would have used the Y5 materials when not required to do so.

The relationship between the trial and 2017 KS2 mathematics tests

The 2017 KS2 mathematics test consisted of three papers: Paper 1, an arithmetic test (maximum 30 marks); and Papers 2 and 3, reasoning papers (maximum 35 marks on each paper). The arithmetic paper consists of abstracted calculations and so its content is different from the application of mathematics in ScratchMaths contexts. Papers 2 and 3 consist of a mixture of shorter questions similarly requiring a single calculation, items testing other mathematical knowledge, or more involved questions requiring multiple steps to derive a solution. In terms of mathematical content, a review of items in comparison with ScratchMaths mathematical content (see Appendix C) suggests that at most a third of the marks across Paper 2 and 3 have content that relates in some way to ScratchMaths module content, and approximately half of that proportion arguably has a close match. Further, some of the questions with the closest match were part of module 6 which focused on coordinates and geometry. Module 6 had the lowest level of implementation (56% of Wave 1 Y6 survey respondents had **not** used any of the module 6 materials).

However, concern about the match between test content and ScratchMaths content only stands if the alternative programme theory of change is posited. Such a revision would view the potential causal relationships between ScratchMaths and mathematical attainment to be due to ScratchMaths being an environment or context for practising and applying mathematical knowledge learnt elsewhere. In the programme theory of change as posited by the ScratchMaths team during the intervention, it was computing and computational thinking that was proposed as leading to a change in mathematical thinking through teacher mediation.

Computational thinking measure

As discussed in the section on background evidence, the construct of computational thinking is a contested one. Scholars' definitions of computational thinking and the means to assess it have yet to be fully agreed. The computational thinking test was developed specifically for this intervention and outcomes on this measure should be treated with caution. As discussed above (and in Appendix F), the ScratchMaths team raised concerns about the appropriateness of the particular questions selected for the computational thinking assessment and its suitability for the trial.

Further, the requirements of a short test with (in most cases) multiple choice items means that the test focused on the unistructural level (Meerbaum-Salent et al., 2013) of computational thinking. The Y5 ScratchMaths materials provide opportunities to develop and/or necessitate the development of multistructural and relational knowledge and skills. It may be that developing multistructural and relational computational thinking skills should entail the development also at the unistructural level. However, this has not been established in previous research or in the current project.

Further, only a limited range of computational thinking concepts were implicated in test items. In particular, understanding or application of concepts of loops, logical expressions and data storing (Brennan and Resnick, 2012) were not tested. Other important concepts such as parallelism and conditionals, which are important aspects of ScratchMaths Y5 materials, were tested on a limited number of items. Considering ScratchMaths as a mathematics intervention, it is also important to highlight the concept of a latent variable, which is an important aspect of Y5 ScratchMaths materials, but not tested in the CT test. It may be that ScratchMaths had a positive impact on aspects of computational thinking that are not tested by the CT test and the positive effect found does not fully capture this.

It is also important to recognise that the CT test measures elements of computational thinking, but it is not a test of computing knowledge or skills or an assessment of attainment on the computing national curriculum. So, the outcome in relation to the CT test cannot be used to infer possible impact of engagement in ScratchMaths on either general computing knowledge or skills or computing curriculum attainment. However, given that there was a difference between control and intervention schools in CT scores, this does suggest that the CT test is measuring at least some aspects of what was being taught as part of the ScratchMaths curriculum. Also, as noted, teachers did consider the intervention effective for learning Scratch programming.

A further potential limitation of the CT test is that some of the metrics from the Rasch analyses of the main data question the validity of the argument that this is a unidimensional scale. As indicated above, the Andersen Likelihood Ratio test on these data was significant. This is indicative that the scale is not unidimensional. However, the main data analyses involved a very large sample of 3964 pupils. It has been argued by some that such large samples are likely to lead to irrelevant model deviations being statistically significant and, as such, an argument could still be made for the unidimensionality of the measure. Further limitations of the CT test are discussed in Appendix F.

School level imbalance for most recent OFSTED (overall effectiveness) inspection

Table 9 compares the baseline sample of intervention and control schools across 12 school level and four pupil level variables. At the pupil level, the intervention and control samples were very similar in terms of FSM, gender and KS1 attainment (overall and in maths). At the school level, the two samples were similar across 9³¹ of the 12 variables. Some imbalance was observed for three variables: School size: on average, intervention schools were smaller (n=397) compared with control schools (n=426); school level %EAL: the mean %EAL was lower for intervention schools (24%) compared with control schools (30%); last OFSTED inspection; a smaller proportion of intervention schools were classed as 'outstanding' (17%) compared with control schools (35%). A higher proportion of intervention schools were classed as good (66%) compared with control schools (62%) and a higher proportion of intervention schools were classed as 'requires improvement' (17%) compared with control schools (4%).

Given the excellent balance across 13 variables, we do not think that the imbalance shown with the subjective OFSTED inspection outcomes serve to undermine the robustness of the trial. The overall effectiveness OFSTED ratings are based on the last inspection and were available for 105 of the 110 schools in the trial. It should be noted that the dates for the last OFSTED inspection ranged between Jan 2007 and June 2015. We included the OFSTED ratings on request from EEF but, given their subjective nature and wide range of dates, we do not think that they provide a valid or useful way of comparing the two samples at baseline. Given how strikingly balanced the two samples are across all measures of attainment at both school and pupil levels for the 2014 academic year, we do not think that the imbalance relating to OFSTED is a problem. Similarly, the slightly smaller schools and lower

³¹ School level KS1 average attainment (2014); KS1 to KS2 maths value added (2014); % with KS2 maths level 5+ (2014); %FSM; %SEN; % Female; Aggregated KS1 maths points score (2014), Aggregated KS1 APS and type of school.

concentrations of EAL pupils in intervention schools compared with control schools do not serve to undermine confidence in the trial and impact evaluation findings

Bias and reliability in implementation and process evaluation data

In Appendix J, details of the process evaluation interview and survey samples are given. Schools that did not achieve sustained participation are under-represented in the sample. Thus, reasons for low fidelity are largely inferred rather than fully established. Survey data were shared with the ScratchMaths team and participants were aware this was the case. This may therefore also have led to bias in responses. However, there is broad agreement between interview data and survey data findings, and the identities of the interview sample were not shared with the ScratchMaths team. Lastly, participants were assured that steps would be taken to ensure anonymity as far as possible.

As noted, when examined at teacher level, there were differences between IoE attendance data and attendance data collected through the Y5 and Y6 Wave 1, intervention surveys... This suggests that fidelity data from the survey about other matters may also not be reliable - for example, the number of modules taught.

On-treatment analyses

The on-treatment analyses used data from 27 of the 55 intervention schools where teachers completed surveys in both Y5 and Y6. Whilst it seems likely that schools which engaged well with ScratchMaths are more likely to have responded to the IPE surveys, having more complete data to draw on would have improved the robustness of the on-treatment analyses.

Further, whilst the ScratchMaths programme was aimed at teachers, fidelity to ScratchMaths was measured at the school level. For example, where we had a response from one teacher, this was used to measure fidelity to ScratchMaths at the school level; where we had a response from two teachers, a single school-level fidelity measure was derived (usually by taking an average, see Appendix J). It would have been preferable to be able to link specific teachers to maths classes (and specific pupils) and then to have undertaken the fidelity analyses at the teacher level. We did collect baseline data that enabled us to identify pupils within maths classes (and therefore include a class level within the analyses). However, sufficient detail was not collected on the specific teacher who was attached to each of these maths classes to enable both class and teacher levels in the analyses.

Generalisability

The trial design focused on schools with two-form entry and these comprised 78% of schools in the trial. It cannot necessarily be assumed that similar outcomes would be found in a sample comprising schools of other sizes. Similarly, in terms of the sample, schools were recruited through Naace and it may be that schools already engaged with computing and Scratch were over-represented in the sample.

Future research and publications

A number of further research questions arise from the evaluation. Teachers considered the activities to be effective in developing computing skills and knowledge, but this was not directly measured in this trial or compared to alternatives, and this area could be addressed in future research. The trial results point to the need for more research to understand reasons for and patterns in the association between computational thinking, computing knowledge and skills, and mathematics attainment, as well as the development of these phenomena.

The ScratchMaths team intend to publish papers on the project design and exploring pupils' views of programming, the ScratchMaths didactical/pedagogical approach and effective implementation, case studies of ScratchMaths implementation, and constructions of fidelity.

The SHU evaluation team intend to publish a paper reporting the overall evaluation of the trial; a report on the development and properties of the computational thinking measure and its relationship to mathematics and English attainment and the implications for this of the way computational thinking is defined; and data from the trial may inform publications drawing on other studies of challenges to engagement and participation in professional and curriculum development for schools.

References

- ACME (2016). *Professional learning for all teachers of mathematics: Principles for teachers, senior leaders and those who commission and provide professional learning*. <http://www.acme-uk.org/media/36491/professional%20learning%20for%20all%20teachers%20of%20mathematics%20-%20final.pdf>
- Alexandrowicz, R. W. and Draxler, C. (2016). Testing the Rasch model with the conditional likelihood ratio test: sample size requirements and bootstrap algorithms. *Journal of Statistical Distributions and Applications*, 3(1), 2.
- Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., ... Wittrock, M. C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives*. New York, NY: Addison-Wesley Longman.
- Back, J., Hirst, C., De Geest, E., Joubert, M. and Sutherland, R. (2009). *Final report: Researching effective CPD in mathematics education (RECME)*. NCETM.
- Benton, L., Hoyles, C., Kalas, I. and Noss, R. (2017). Bridging Primary Programming and Mathematics: some findings of design research in England. *Digital Experiences in Mathematics Education*, pp. 1-24.
- Benton, L., Hoyles, C., Kalas, I. and Noss, R. (2016) *Building mathematical knowledge with programming: insights from the ScratchMaths project*. Constructionism, Bangkok, Thailand, 02 Feb 2016 - 05 Feb 2016. 02 Feb 2016 (Conference proceeding)
- Blakemore, L. (2017). *Does teaching computer programming within Key Stage 1 of the primary curriculum enhance children's problem solving skills?* (Doctoral dissertation, University of Sheffield).
- Bond, T.G. and Fox, C.M. (2015). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd Edition)*. Routledge.
- Brennan, K., & Resnick, M. (2012, April). New frameworks for studying and assessing the development of computational thinking. In *Proceedings of the 2012 annual meeting of the American Educational Research Association*, Vancouver, Canada (pp. 1-25).
- Clements, D. (2000). From exercises and tasks to problems and projects unique contributions of computers to innovative mathematics education. *Journal of Mathematical Behavior*, 19, 9-47.
- Cohen, J., Cohen, P., & Stephen, G. (2003). West, and Leona S. Aiken. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ.
- Computational thinking with Scratch (n.d.) *How do I assess the development of CT?* Retrieved Sept 2017 <http://scratched.gse.harvard.edu/ct/assessing.html>
- Cordingley, P., Higgins, S., Greany, T., Buckler, N., Coles-Jordan, D., Crisp, B. and Coe, R. (2015). *Developing great teaching: lessons from the international reviews into effective professional development*. Teacher Development Trust. <http://dro.dur.ac.uk/15834/1/15834.pdf>
- Cuny, J., Snyder, L., & Wing, J. M. (2010). *Demystifying computational thinking for non-computer scientists*. Unpublished manuscript in progress, referenced in <http://www.cs.cmu.edu/~CompThink/resources/TheLinkWing.pdf>.
- DfE (2013). *Computing programmes of study: key stages 1 and 2. National curriculum in England*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/239033/PRIMARY_national_curriculum_-_Computing.pdf

DfE (2014). *Mathematics programmes of study: Key Stages 1 and 2*. London: DfE. Url: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/335158/PRIMARY_national_curriculum_-_Mathematics_220714.pdf October, 2014.

Fuller, U., Johnson, C. G., Ahoniemi, T., Cukierman, D., Hernán-Losada, I., Jackova, J., ... & Thompson, E. (2007). Developing a computer science-specific learning taxonomy. *ACM SIGCSE Bulletin*, 39(4), 152-170.

Furber, S. (2012). *Shut down or restart? The way forward for computing in UK schools*. The Royal Society.

Gadernann, A. M., Guhn, M. and Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation*, 17(3), 1-13.

Holmes, W.M. (2014). *Using Propensity Scores in Quasi-Experimental Designs*. Sage.

Hoyles, C., & Noss, R. (1992). *Learning mathematics and Logo*. Cambridge: MIT Press.

ISTE and CSTA (nd) *Operational Definition of Computational Thinking for K–12 Education*. <https://csta.acm.org/Curriculum/sub/CurrFiles/CompThinkingFlyer.pdf>

Lee, W. C. (1990). *The effectiveness of computer-assisted instruction and computer programming in elementary and secondary mathematics: A meta-analysis*. PhD dissertation, University of Massachusetts (January 1, 1990). Electronic Doctoral Dissertations for UMass Amherst. Paper AAI9022709.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions*, 16 (2), 878.

Mair, P., Hatzinger, R., Maier M. J. & Rusch, T., (2015). *Package ‘eRm’. Version 0.14-0*.

McMaster, K., Rague, B. and Anderson, N. (2010). Integrating mathematical thinking, abstract thinking, and computational thinking. Paper presented at *ASEE/IEEE Frontiers in Education Conference*, October 2010, Washington DC.

Meerbaum-Salant, O., Armoni, M., & Ben-Ari, M. (2013). Learning computer science concepts with scratch. *Computer Science Education*, 23(3), 239-264.

Monroy-Hernandez, A. and Resnick, M. (2008). Empowering kids to create and share programmable media. *Interactions*, March and April 2008, 50-53.

Naace (n.d.) *Naace professional development standards*. <https://www.naace.co.uk/pd/standards/>

Opfer, V. D. and Pedder, D. (2011). Conceptualizing teacher professional learning. *Review of Educational Research*, 81, 376-407.

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests*.

Resnick, M., Maloney, J., Monroy-Hernández, A., Rusk, N., Eastmond, E., Brennan, K., ... & Kafai, Y. (2009). Scratch: programming for all. *Communications of the ACM*, 52(11), 60-67.

Selby, C. and Woollard, J. (2013). *Computational thinking the developing definition*. http://eprints.soton.ac.uk/356481/7/Selby_Woollard_bg_soton_eprints.pdf

Selby, C., Dorling, M. and Woollard, J. (2014). Evidence of assessing computational thinking. *Authors' original*. URL <http://eprints.soton.ac.uk/372409/1/372409EvidAssessCT.pdf>

Stoll, L., Harris, A. and Handsome, G. (2012). *Great Professional Development that leads to great pedagogy: nine claims from the research*. NCTL.

Straw, S., Bamford, S. and Styles, B. (2017). *Randomised Controlled Trial and Process Evaluation of Code Clubs*. Slough: NFER.

van Driel, J. H., Meirink, J. A., van Veen, K., & Zwart, R. C. (2012). Current trends and missing links in studies on teacher professional development in science education: a review of design features and quality of research. *Studies in science education*, 48(2), 129-160.

Wing J. M. (2011) *Research Notebook: Computational Thinking - What and Why?* Retrieved from <http://scratched.gse.harvard.edu/ct/defining.html>

Wing, J. M. (2008). Computational thinking and thinking about computing. *Philosophical transactions of the Royal Society of London A: mathematical, physical and engineering sciences*, 366 (1881), 3717-3725.

Wright, B.D., and Stone, M.H. (1999). *Measurement essentials*. Wide Range Inc., Wilmington.

Yoon, K. S., Duncan, T., Lee, S. W. Y., Scarloss, B., & Shapley, K. L. (2007). *Reviewing the Evidence on How Teacher Professional Development Affects Student Achievement*. *Issues & Answers*. REL 2007-No. 033. Regional Educational Laboratory Southwest (NJ1).

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.

ScratchMaths is a **very low-cost** intervention when delivered by the developers and would be so if delivered by local hubs.

Appendix B: Security classification of trial findings

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%	5		5	
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%		Adjustment for Balance [x]		
3	Well-matched comparison (using propensity score matching, or similar) or experimental design with moderate concerns about validity	MDES < 0.4	21-30%			Adjustment for threats to internal validity [x]	
2	Weakly matched comparison or experimental design with major flaws	MDES < 0.5	31-40%				
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 = 5
- **Reason for adjustment for balance** (if made): N/A
- **Reason for adjustment for threats to validity** (if made): N/A
- **Final padlock score:** initial score adjusted for balance and internal validity = 5

Appendix C: ScratchMaths content and ScratchMaths team theory of change

The figures below provide detailed lists of Scratch commands and concepts, computing concepts and mathematical content covered in ScratchMaths. These lists were derived from information on ScratchMaths materials.

Scratch concepts and commands

Figure 7: ScratchMaths concepts and commands

Sprite	Backdrop	If ... Then ... block
Stage	Pre-defined blocks	... < ..., ... > ... blocks
Block	Pick random	Broadcast blocks
Stamp block	Repeat block	Say... blocks
Hat block	Define block	Costume # block
Turn block	When this sprite clicked block	... + ..., ... - ..., ... = ...blocks
Snapping blocks	Hide and Show blocks	Stop all block
Script Move block	Graphic effects block	... of ... block
Repeat block	Change by ... and Set to ... blocks	Ask and answer blocks
Costume	Forever block	Join block
Define block	If on edge, bounce block	... * ..., ... / ... blocks
Pen down, pen up blocks	Point towards... block	<variable name> block
Pen colour blocks	Repeat Until..., Touching... blocks	Set <variable name> to ... block
Pen shade blocks		When ... key pressed
Pen size blocks		

Computing concepts

A number of computing concepts were addressed in more than one module, and in some cases this was specified in the core content of modules. However, once a concept was introduced it was then employed multiple times and this may not have always been highlighted within the module content list if it was not a key learning focus of the activity

Figure 8: ScratchMaths computing concepts

Algorithm	Initialisation
Broadcasting and receiving messages	Logical reasoning
Broadcasting conditions	Multiple actors
Command	Multiple costumes and animation
Conditions and conditional loops	Parallel behaviours
Debugging	Repetition
Decomposition	Sequence
Definitions	Selection
Events	Variable
Expressions	

Mathematical content

Figure 9 below provides the mathematical content identified on ScratchMaths investigations. Numbers in brackets indicates they were listed on the summary description at the start of more than one investigation. As stated in the intervention description, each module consisted of a number of investigations, with an investigation consists of core activities that have certain steps designated as extensions, as well as some further, separate extension activities.

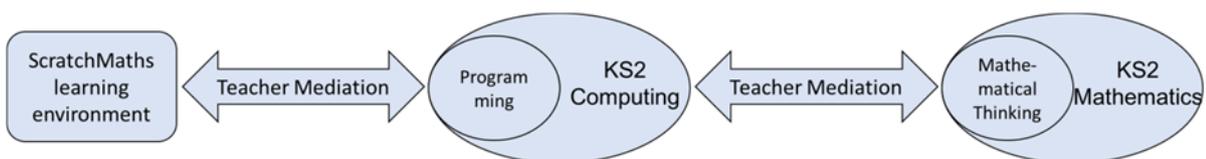
Figure 9: Mathematical content

Number	Geometry	Measurement	Algebra (Y6 only)	Cross strand
Addition, subtraction Multiplication and division (3) Factor pairs/Factors (2) Place value Roman numerals Positive and negative numbers (3) Division & rounding Fractions	Angles (4) Coordinates (5) Reflection Regular and irregular polygons (3) Rotation (2) Scale Factor Symmetry Transformation Translation (2)	Conversions Estimations Perimeter (2) Time, Weight, Length	Algebraic expressions Patterns Sequences (2)	Mathematical modelling Problem solving
			Ratio and proportion (Y6 only)	Not included explicitly
			Ratio and proportion	Random numbers (2) Randomness (2)

ScratchMaths team theory of change

As noted on page 6, during the third year of the trial the ScratchMaths team proposed an alternative to the theory of change agreed at the start of trial and presented in the evaluation protocol (see figure 1). This is presented in figure 10 below.

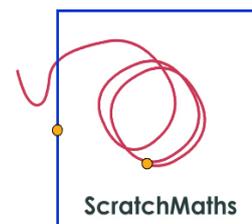
Figure 10: ScratchMaths team 2017 theory of change



Appendix D: Consent forms and MoUs

D1 Parental information and Opt-out consent form

ScratchMaths project information for parents/carers and opt-out form



Dear parent/carer,

Your child's school is taking part in a trial project designed to improve mathematics teaching and learning. The particular focus is on computational thinking, a problem solving method that uses computer science techniques (http://en.wikipedia.org/wiki/Computational_thinking). The idea of using computing to support maths learning has been shown to work over many years but never tested at scale with large numbers of children.

The project is being run from the London Knowledge Lab (LKL), UCL Institute of Education and supported by the Education Endowment Foundation (EEF). Project partners include Sheffield Hallam University (SHU), London Connected Learning Centre (CLC) and The National Association for the Advancement of Computer Education (Naace). The project leaders Professor Richard Noss, Professor Dame Celia Hoyles and Professor Ivan Kalas have many years' experience of running similar projects.

The project will develop lessons using the Massachusetts Institute of Technology (MIT)'s Scratch online software – which teaches children the building blocks of computer programming. The project will test what effects learning with our materials based on Scratch have on pupils' learning of mathematics. There will be two sets of lessons, one on computational thinking and one on mathematical thinking. Teachers will deliver the new lessons to Year 5 and 6 classes.

Schools will take part in one of two ways. In some schools (called treatment schools) teachers will take part in professional development activities and will be provided with new materials to teach ScratchMaths in summer/autumn 2015. The professional development activities will be undertaken by Naace and the University of London team. In other schools (called control schools) teachers will carry on as normal until summer/autumn 2016 when they will have CPD and start teaching with the new materials.

By comparing the two sets of schools we will be able to judge if, and how the approach works. This has the potential to benefit not just your child but children in schools across the country. (The 100 schools will be randomly split into two groups of 50 in March 2016).

As part of this evaluation we need 2 classes of pupils in Year 5 in your child's school to take an extra 30-minute computational thinking test (administered by SHU) in Summer 2016. The tests taken are appropriate to pupils' ages. Pupils will be told a few weeks before if they have been chosen to take the extra test. The treatment group schools will also be taught the national curriculum aligned lessons as detailed above as part of their maths and computing periods at school.

The data collected in this project will be used to inform reports and publications about the project. No individual child will be identified and all data are stored securely to ensure compliance with the 1998 Data Protection Act. We will anonymise your child's name by using an identifier number in security protected computer files. The Department for Education may use the identifier number in the future to link data from this project to data that is routinely collected on pupils by the government, for example exam results.

If you would prefer your child NOT to take part in the additional one hour test, please complete the slip overleaf and give it to your child's teacher/ take it to the school office. If you would like more information, please contact us on 020 7911 5577 or j.otoole@ioe.ac.uk

Please return this slip to your child's teacher (or school office)

if you DO NOT wish your child to be involved in

ScratchMaths project evaluation

I DO NOT give my permission for my child to take part in the ScratchMaths project evaluation and testing.

Child's full name	
Signed	
Parent/carer	
Date	

D2 Memorandum of understanding

The ScratchMaths project is a randomised control trial that seeks to establish if the learning of computer programming in Scratch can improve not only computational thinking but also mathematics performance at Key Stage 2. It runs from late 2014 to the middle of 2017 and if successful, will subsequently be rolled out across the country. Please read through the following information and sign and return one copy of this document if you wish to join the project.

We hope that your school will take the opportunity to be one of 100 schools in England to participate in this innovative project based at the Institute of Education, University College London and supported by the Education Endowment Foundation (EEF). Project partners include Sheffield Hallam University, London Connected Learning Centre (CLC) and The National Association for the Advancement of Computer Education (Naace).

This project will produce materials and offer professional development that is aligned with the Computing Curriculum (Y5) and the Mathematics Curriculum (Ys 5 and 6). It will aim to boost mathematics scores at KS2 by approaching some of the mathematics involved through creative programming in Scratch.

Participating schools will normally be two-form entry. Schools will also need to meet the technical requirements set out in the memorandum of understanding.

The project is a randomised control trial – similar to those used in medicine to test the effectiveness of a new treatment. The 100+ schools selected will be randomly split into 50 *treatment schools* and 50 *control schools* in March 2015.

Involvement in the CPD and delivery of the interventions will be staggered.

Treatment schools will

- receive two specially designed curriculum-aligned interventions for Ys 5 and 6 in *computational thinking* and *mathematical thinking*, which include free student materials and teacher guidance using MIT's Scratch software (<http://scratch.mit.edu/>). In addition we will provide two days of CPD, separated by a couple of weeks, in each of the summer terms 2015 and 2016 after KS2 testing.
- receive free CPD for teachers who will be teaching the interventions (computational thinking in Y5 and ScratchMaths in Y6).
- be invited to participate in an online teacher community for mutual support and advice. be require not to share Y5 project resources with other schools (in year one of the project) and Y6 project resources (in year two)

Control schools will receive free access to Y5 ScratchMaths materials and training in computational thinking in summer 2016. They will receive all of the materials for Y6, in 2017. They will be invited to participate in an online community in summer 2016.

All schools will receive feedback on the outcomes of the study to inform future practice.

Sheffield Hallam University will be conducting the analysis of the effects of the intervention. To take part, schools will need to provide Sheffield Hallam with data on teachers who will be taking part as well as the Unique Pupil Numbers (UPNs) for all Year 5 pupils in the school. Schools will also need to inform parents and give them the choice for their children to opt-out.

To see if the programme is effective, Sheffield Hallam University will retrieve KS1 scores for the Y5 pupils from the National Pupil Database. In Summer 2016, the Y5 pupils will take an on-line test of computational thinking arranged by Sheffield Hallam. Support will be given to schools to administer this. Test taking can be staggered and schools will be able to choose when the pupils will take the test, within a given time period. At the end of the trial, Sheffield Hallam will retrieve KS2 data from the National Pupil Database.

	Treatment schools	Control schools
Application to apply	Complete MOU. Provide pupils UPN to SHU	Complete MOU. Provide pupils UPN to SHU
Summer '15	CPD computational thinking (Y5)	(ongoing school activities for Y5 computing)
Autumn '15 /Spring '16	Computational thinking intervention (Y5) online teacher survey	(ongoing school activities for Y5 computing) online teacher survey
Summer '16	CPD ScratchMaths Computational thinking test (Y5) online teacher survey	CPD computational thinking Computational thinking test Y5 online teacher survey
Autumn '16 /Spring '17	CPD ScratchMaths intervention (Y6) online teacher survey	Computational thinking intervention (Y5) online teacher survey Computational thinking test (Y5)
Summer '17	Key Stage 2 Mathematics test as normal	Key Stage 2 Mathematics test as normal

100 schools will be selected for the trials and randomly split into 50 treatment schools and 50 control schools. Prior to the trial starting ALL schools will...

Technical requirements

- ensure adequate Internet connectivity is available for Y5 and Y6 pupils at the time of the interventions.
- ensure enough machines are available (one between two at least) for the Y5 and Y6 pupils at the time of the interventions.
- to run Scratch 2.0 online with the whole group/class of pupils in parallel, you will need a relatively recent web browser (Safari, Chrome 7 or later, Firefox 4 or later, or Internet Explorer 8 or later) with Adobe Flash Player version 10.2 or later installed. Scratch 2 is designed to support screen sizes 1024 x 768 or larger.
- Scratch 2.0 does not work on iPads and similar devices, so we advise testing Scratch on the computers to be used before agreeing to take part in the project.

NOTE: The project team has received written consent from MIT's Scratch team that each school will be allowed to set up individual Scratch accounts for each of their participating pupils.

Data

- provide information on request about the school and two teachers who will be involved in the project. This includes teachers' attendance at CPD, current activities related to programming and the National Curriculum for Computing; information on any other use of Scratch programming in the school; information on any testing procedures for computing. A link to a short online form will be sent to each school for completion.
- issue information about the project and opt - out consent forms to parents and provide details of any parents opting out.

- Provide Unique Pupil Numbers (UPNs) for the pupils who are involved in the trial and agree that the research team can access the National Pupil Database to retrieve KS1 and KS2 data as well as demographic data such as free school meals status, gender and so on.
(no individual school or pupil will be named in any report or publication arising from the research.)

Surveys/testing

- agree for participating teachers to take part in online surveys during 2015/17.
- allow pupils to take part in the Y5 computational thinking test in Summer 2016, administered by SHU (30 minutes maximum length), and for members of the Sheffield Hallam team to visit if needed to review how the tests are conducted.
- agree to record any issues that might affect the fidelity of the implementation of the intervention e.g. any changes in the teachers who deliver the interventions.

Schools selected as TREATMENT SCHOOLS will...

Teaching

- allow for up to 20 hours of teaching time per school year for engagement with the specially designed curriculum units (computational thinking, Y5, 2015-16 leading to ScratchMaths, Y6, 2016-2017).

CPD

- allow two teachers to attend CPD sessions (after KS tests), two days Summer 2015 and two days Summer 2016. The teachers trained will be those who are teaching the targeted group the following September, (that is Y5 for September 2015/16 and Y6 for 2016/17).

The research

- allow members of the ScratchMaths project team to visit and observe lessons at pre-arranged convenient times.
- allow participating staff to take part in, for example, research interviews, surveys and events as required by the project within reasonable scope of their time and availability.
- **Materials** be require not to share Y5 project resources with other schools (in year one of the project) and Y6 project resources (in year two)

Schools selected as CONTROL SCHOOLS will...

Teaching

- continue your normal teaching programme of computing and mathematics 2015/16
- choose to use intervention computational thinking materials in 2016/17

CPD

- allow two teachers to attend CPD sessions, two days in Summer 2016.

The research

- allow members of the ScratchMaths project team to visit and observe lessons at pre-arranged convenient times.
- allow participating staff to take part in research interviews, surveys and events as required by the project within reasonable scope of their time and availability.

The PROJECT will make the following commitments: *Feedback*

- provide feedback on how the school could develop its approach to pupils' development of computational thinking by providing responses to computational thinking tests.

CPD and student/teacher materials

- provide professional development and all the materials for Y5 and Y6 as set out above by the end of the project.

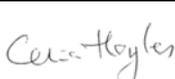
Ethics/ Data protection

- perform all necessary ethical checks to make sure school staff and pupils and project researchers are acting in accordance with ethical procedure. All researchers entering school premises will hold a current DBS (formerly CRB) check.
- Test results and pupil data will be treated with the strictest confidence. The tests results will be matched with data from the National Pupil Database(including if available item-by-item scores) and potentially other government data sets, and shared with researchers at Sheffield Hallam University, The University of London, the Education Endowment Foundation's data archive and the UK data archive for research purposes We will not use pupil name or the name of the school in any report arising from the research.
- We will provide outcomes of the tests to your child's school so that the teachers can use the results to decide how they can help children to learn more
- ScratchMaths project resources are developed by the London Knowledge Lab, UCL Institute of Education under a grant from the Education Endowment Foundation. Unless stated otherwise, LKL is the owner of all intellectual property rights in the materials. These works are protected by copyright laws and treaties around the world. All such rights are reserved.

Support for computational thinking test

- Sheffield Hallam University will provide written, email and telephone support for testing

If the above terms are acceptable, please sign and date both copies, keeping one copy for your records and returning the other to James O'Toole, ScratchMaths project, LKL, 23 Emerald St, London, WC1N 3QS or by email to j.otoole@ioe.ac.uk

Signed		Signed	
Name	Prof. Dame Celia Hoyles, ScratchMaths project	Name	Professor Richard Noss, ScratchMaths project
Signed			
Name	Dr Mark Boylan, Evaluator, Sheffield Hallam University		

Name		School	
Role	Headteacher	Date	

Appendix E: Detail of team roles

ScratchMaths development team, University College London, Institute of Education (IoE)

Professor Richard Noss: Principal Investigator

Overall Project Investigator with responsibility for delivery of all outcomes including: overseeing project to ensure recruitment; quality of provision and implementation; consistency of aims and methods for computational thinking and for mathematics in the CPD; ensuring the design research and work in schools provides evidence of the key pedagogical and teacher factors that underlie any success.

Professor Celia Hoyles: Co-Principal Investigator

Particular responsibility for mathematics CPD; qualitative outcomes and reports; managing the (complex) logistics of training in hubs; outward-facing liaison with schools, teachers; dissemination in general.

Professor Ivan Kalas: Co-Investigator

Responsible for literature review (in particular Beaver texts published in Slovak or Czech languages); responsible for design and drafting all Scratch tasks, building a selection of Beaver tasks (from previous years' competitions) and extending this selection by several Beaver-like tasks; co-ordination of design evaluation.

Professor Dave Pratt: Co-Investigator

Responsible for final design and validation of Scratch tasks (and design of maths Scratch tasks, both with Ivan Kalas); aligning the design evaluation and the process evaluation, in relation to fidelity measures and school-level data collection; drafting qualitative reports of first intervention.

Research Officer: Dr Laura Benton

Responsible for day-to-day liaison with schools; data collection; analysis under supervision; implementation of training as designed for treatment and control; alignment of tasks and activities with school curricula and practice; assistance to teachers in CPD sessions for technical and organisational issues.

Research Officer: Piers Saunders

Responsible for design and validation of maths Scratch tasks and resources with Ivan Kalas. Responsible for design and delivery of ScratchMaths CPD with Laura Benton and Ivan Kalas.

Dr Alison Clark-Wilson

Responsible for day-to-day liaison with schools; data collection and analysis and dissemination

Project Administrator: James O'Toole

Responsible for maintenance of databases; design of website and maintenance; organisation of school visits; liaison with external bodies; coordination of staffing.

IOE PhD students: Piers Saunders, Johanna Carvajal

Mark Chambers: CEO

Responsible for recruitment and leading the design of training and the training team (for all treatment teachers and for control teachers).

Advisory Group members

Professor Janet Ainley, University of Leicester
Miles Berry, University of Roehampton
Joe Halloran, Lambeth City Learning Centre
Gillian Ingram, Manager, Camden City Learning Centre
Debbie Morgan, Director for Primary, National Centre for Excellence in the Teaching of Mathematics Dr.
Mary Webb, King's College London.

Evaluation Team, Sheffield Hallam University (SHU)

Professor Mark Boylan: Evaluation project director / Principal Investigator
Sean Demack: Lead statistician
Dr John Reidy: CT test design
Anna Stevens: Data management, analysis of pupil progress, trial management of the CT test
Claire Wolstenholme: Project manager

Sarah Reaney-Wood: Research Fellow, data management, fidelity analysis
Dr Martin Culliney: Research Fellow, statistical analysis support.
Ian Guest: Research associate, process evaluator
Professor Hilary Povey: Programming in Mathematics advisor and process evaluator
Phil Spencer: Computer Science in Primary advisor

Ian Chesters: Administrator

Appendix F: Computational thinking test - development and analysis

CT test - introduction

This appendix describes the development of the tests of aspects of computational thinking used in the ScratchMaths evaluation, as an intermediate and secondary measure. The test was developed independently by Sheffield Hallam University. Prior to development of ScratchMaths materials, the ScratchMaths team supported the initial stages of development by informing SHU of their operational definition of computational thinking, as used in the intervention design, and also discussed examples of Bebras/Beaver questions.

Aim and design criteria

Aim

The CT test aimed to provide an intermediate and secondary measure for the ScratchMaths trial to assess the mechanism of change by which the ScratchMaths intervention was proposed to improve mathematics attainment. The original underlying theory of change supposed that changes in KS2 attainment attributable to the ScratchMaths intervention will arise from enhancement of computational thinking that is related to mathematical thinking.

Design criteria

Criteria that informed design were:

1. To avoid a test that was inherent to treatment, that is to ensure the test was not too closely aligned to ScratchMaths materials or tasks, for example by testing knowledge of Scratch syntax.
2. Delivered online and taking no longer than 30 minutes, and scored by coding or algorithm.
3. Using or adapting existing problems/test items that were recognised as being related to computational thinking thus providing construct validity.
4. Items should relate to constructs of computational thinking in research literature.
5. Items should cover a range of different computational thinking skills/capacities.
6. Accessible for the full cohort of Y5 with exception of SEN students (as determined by the school) who would usually be dis-applied from mathematics tests using Beaver challenges designed for this age of pupils as a model
7. Provide a spread of scores across 10 items.

Overview of the CT test development process

Overview

Initial scoping: Members of the SHU and IoE teams met in November 2014 to develop a shared understanding of IoE's operational definition of computational thinking as used in the intervention design.

Following this, SHU designed, developed and tested the CT measure independently. The CT test design was undertaken in four steps:

- Initial design review of Beaver/Bebras³² questions and other items by SHU, leading to a first draft of the scale.

³² Established in 2004, see <http://www.bebbras.org> or <http://www.beaver-comp.org.uk> or <http://www.ibobor.sk/> it is now run in 30 countries. In 2013, more than 720,000 pupils took part in the contest.

- Trial of test items in two schools (*planned* - 60 Y5, 60 Y6 pupils, 120 in total; *actual* – 49 Y5 pupils, 66 Y6 pupils, 115 in total) and descriptive statistical analysis of items/scale, leading to a revised scale and test protocols.
- Test of the revised scale plus focus on trial of protocols in one school. The plan was initially for 30 Y5 and 30 Y6 pupils, but in the event, due to school availability, it was trialled with 57 Y6 pupils. However, the purpose here was to trial the revised test protocols, with the teachers administering the test independently, and so the age of the pupils was not considered relevant.
- Full test to establish the correlation between CT test scores and both KS1 and KS2 assessment scores with six schools (*planned* - 360 Y6 pupils; *actual* - 231 [Y6] pupils for which Unique Pupil Numbers (UPNs) and opt-out consent were obtained and matched to the National Pupil Database (NPD)). The initial intended number was based on a model of a moderate correlation (0.3) and a criterion for significance of .05.

Design recruitment and ethics

There were nine schools in total involved. The schools were recruited in the Sheffield City region via a notice in the local maths hub newsletter and from direct contact with schools that are part of SHU's research-engaged practitioners' network. The schools were not involved in either the IoE design process or the main trial. Each school in the CT test design process was offered the opportunity for two teachers to go to a one-day and a separate half-day professional development session run by SHU on teaching computing in primary schools. Ethics and consent followed the same procedures as the main trial. Parental opt-out consent forms were issued by schools to all parents of pupils in classes who would be taking the test. Forms were then collected by the class teacher and any pupils whose parents had returned the form did not take the test. Opt-out forms were managed by the class teachers and SHU were not notified of opt-outs.

Computational thinking³³

In the main body of this report (pages 9-11) computational thinking is considered. The discussion that follows develops this in relation to the way in which computational thinking is viewed within the Scratch community, and this informed CT test development. Brennan and Resnick (2012) offer the following framework:

Figure 11: Scratch-informed framework for computational thinking

<p>Computational concepts</p> <p>Sequence: identifying a series of steps for a task. Loops: running the same sequence multiple times. Parallelism: making things happen at the same time. Events: one thing causing another thing to happen. Conditionals: making decisions based on conditions. Operators: support for mathematical and logical expressions. Data storing: retrieving and updating values.</p> <p>Computational practices</p> <p>Experimenting and iterating: developing a little bit, then trying it out, then developing more. Testing and debugging: making sure things work – and finding and solving problems when they arise. Reusing and remixing: making something by building on existing projects or ideas. Abstracting and modularizing: exploring connections between the whole and the parts.</p> <p>Computational perspectives</p> <p>Expressing: realising that computation is a medium of creation - “I can create.”</p>

³³ The discussion in this section draws on a presentation by Professor Ivan Kalas shared at a meeting between SHU and the ScratchMaths team in November 2014 in which a number of the cited texts were referenced or quoted from. Following this, further review of these and additional texts was undertaken and the views presented here are the responsibility of the evaluation team and do not represent, necessarily, those of the ScratchMaths team.

Connecting: recognising the power of creating with and for others - "I can do different things when I have access to others."

Questioning: feeling empowered to ask questions about the world - "I can (use computation to) ask questions to make sense of (computational things in) the world."

Building on Brennan and Resnick's emphasis of the importance of creativity as intrinsic to a computational perspective, Kalas (2014) proposes a creative computing perspective in which personal interests, agency and creativity are all important for meaningful engagement in the creation of computational artefacts. Within the ScratchMaths project, this is operationalised through a '5E approach': envisage, explore, exchange, explain, extend.

Fuller et al. (2007) propose a conceptual taxonomy suitable for computer science education that has five categories:

- Prestructural: Mentioning or using unconnected and unorganised bits of information which make no sense.
- Unistructural: A local perspective where mainly one item or aspect is used or emphasized. Others are missing, and no significant connections are made.
- Multistructural: A multi-point perspective, where several relevant items or aspects are used or acknowledged, but significant connections are missing and a whole picture is not yet formed.
- Relational: A holistic perspective in which meta-connections are grasped. The significance of parts with respect to the whole is demonstrated and appreciated.
- Extended abstract: Generalisation and transfer so that the context is seen as one instance of a general case.

Meerbaum-Salant et al. (2013) synthesise the above taxonomy with adaptation of Bloom's taxonomy (Anderson et al., 2001) to propose a two-dimensional taxonomy of three categories adopted from the Fuller et al. (2007) taxonomy - unistructural, multistructural, and relational, each having three sub-levels - understanding, applying and creating.

Assessing computational thinking

Proposals for assessment of computational thinking have been made that seek to do such assessment within and as part of assessment of computing activity and its outcomes. Selby, Dorling and Woollard (2014) map computational thinking components to a pathway to assess progress on the computing curriculum in England. Brennan and Resnick (2012) propose that computational thinking can be assessed through assessment of Scratch design activities including portfolio analysis, artefact-based interviews and design scenarios, among others. This developmental view seeks to account for the complexity and enmeshed nature of different aspects of computational thinking. Meerbaum-Salant et al (2013) designed a pre-test, post-test study specifically to test conceptual understanding within the Scratch context, with items testing understanding of concepts such as repeated execution, variable and event handling. Thus, their test linked conceptual understanding to programming skills. Given the design criteria above, their test was not suitable in the trial as it was specific to Scratch.

Recently, in England, Beaver/Bebras items were used as a measure in an evaluation of Code Clubs (Straw, Bamford, and Styles, 2017). This evaluation made use of a test designed and scored by Chris Roffey of Beaver UK. Some of the items selected for use in the CT test used in the ScratchMaths trial were the same. Chris Roffey provided supplementary detail on the code club test additional to that found in the evaluation report. In the code club evaluation, the test mirrored the usual approach to scoring in Bebras tests. Students were given an initial score of 21 points. The 12 items were designated as A, B or C items with different points awarded for each type of question and points deducted for wrong answers.

Operationalising a definition of computational thinking in the CT test development

Given the limitations of a short, timed, online test, it is clear there are limits to what aspects of computational thinking can be measured. Although some test items require processing and working with multiple sources of information and/or models with more than one step, given the nature of the context, the concepts that are testable are unistructural across the three levels of computational thinking: understanding, applying and creating (Meerbaum-Salant et al, 2013). In the context of the test, in order to get a correct answer, both understanding and applying would be required. In order to give opportunities to 'create' (albeit limited and only possible at the limited unistructural level), it was important to include items that were not multiple choice, but required some construction of a solution by clicking and dragging.

Within this level, test items potentially could assess, or at least indicate the capability to: abstract, decompose, think algorithmically and evaluate (Selby and Woollard, 2013). However, given the limits of the test and the platform, the items did not give opportunities to assess levels of generalisation.

Arguably, both the Meerbaum-Salant and Selby and Woollard frameworks are not specific to computing environments. For example, they are relevant to general mathematical problem-solving or computational maths (McMaster, Rague and Anderson, 2010). However, the Beaver/Bebras and similar items potentially require a range of computational conceptual thinking including: sequencing, proto-loops, parallelism, events and conditionals (Brennan and Resnick, 2012). The term 'proto-loops' refers to running a sequence a small number of times and in a limited context.

Drawing on these various conceptualisations provides a framework for considering the construct validity of the test items for the age group, as summarised in Table 38.

Table 38: Conceptual framework for short online CT test

Unistructural level	Mode of thinking	Concepts
Understanding and applying	abstraction decomposition	sequencing events
Creating	algorithmical evaluation	proto-loops conditionals parallelism

Test development and choice of items

Given the context of a short online test, it is clear that there are many aspects of computational thinking rooted in problem-solving and design over time (as described above) that would not be able to be assessed by the CT test. However, the test aimed to assess elements of computational thinking concepts (Brennan and Resnick, 2012) as well as, in a specified and relatively simple context, types of computational thinking (Selby and Woollard, 2013). Given this limitation, it is more accurate to describe the test as an 'elements of computational thinking' test.

From the agreed understanding of computational thinking, an initial pool of seven potential questions was developed, either taken from previously developed and published puzzles, or generated by the team, to measure the agreed characteristics of computational thinking. This initial bank of questions was then reviewed by the SHU evaluation team for suitability with Y5/6 students. It was agreed that most of these initial items (four) were not appropriate in terms of their match to likely attainment and understanding of the target students.

As a result, the development team drew on tasks developed for an international computing challenge - the Beaver contest³⁴, an international initiative with the goal to promote informatics (computer science, computing) and computational thinking among pupils at primary, lower secondary and upper secondary

³⁴ Established in 2004, see <http://www.bebas.org> or <http://www.beaver-comp.org.uk> or <http://www.ibobor.sk/> it is now run in 30 countries. In 2013, more than 720,000 pupils took part in the contest.

stages. Each year the contest comprises a set of tasks (items) to be completed online, of which some are directly related to computational concepts and computational practices. The use of Beaver items provided a degree of construct validity with items suitable for an online test. Further, Beaver questions are categorised by age and have been used extensively with children of specified ages. This supported identification of age-suitable items. We considered tests from 2013 and 2014 for the three youngest age groups: the Kits (Years 2 and 3), Castors (Years 4 and 5) and Juniors (Years 6 and 7), selecting seven items that appeared on all three test levels in a given year. The aim here was to ensure accessibility across attainment levels. On any individual test item, there are considered to be of different levels of difficulty - A, B, C.

The decisions regarding the number and type of items were taken with the aim of generating an instrument which was age-appropriate, would generally take about 30 minutes to complete, and could be implemented using the Qualtrics survey platform. In relation to age-appropriateness, we aimed for a test that overall had a similar level of challenge to the Castors test. Three members of the SHU team reviewed the Kits and Castors questions and identified items that potentially could test a range of computational thinking concepts/abilities. Readability and other clarity criteria were employed in refining test items.

A 10-item test was generated (see below). This consisted of seven Beaver items from the 2013 and 2014 Beaver tests (*LIFO ice-cream parlour*³⁵, *ice-cream cones*, *water the flowers*, *select a picture*, *in the forest*, *magic bracelet*, and *flip-flop*), one similar to an exemplar Beaver item from 2018 (*park the car*) and two newly-developed items (*supermarket dash*, *colour machine*). The first of these new items was designed to provide an additional question, along with *park the car* which tested aspects of computational thinking in relation to movement (a common programming context for the age group we were concerned with) and was similar to Bebras items on the 2013 and 2014 test which were not suitable due to limitations of the platform used for the test. The other item, *colour machine*, had a similar logical structure to another Beaver item that tested sequencing on the Castors 2013 test (*zebra tunnel*). Here the rationale for modification was to reduce the linguistic demand, but also the number of steps involved in solving the problem. Items were analysed in relation to unistructural levels of computational thinking and mode of thinking. The overall level of challenge of the test was modelled on Beaver items for the age of children taking the test.

These questions were then set up in Qualtrics using the *drag'n'drop*, *multiple-choice* and *ranking options* response formats. The questions were ordered for the initial pilot so that some questions considered to be easier would be encountered towards the start of the test (for example, questions that only appeared on Kits tests). This was in part an ethical decision to ensure that all students would encounter questions they could attempt. Otherwise, questions were sequenced randomly. Following the pilot, given the limited resource for development, the same test order was maintained. In addition to the test questions, two dummy questions were included at the start to ensure that test takers understood the format and the two ways to respond - selecting from a multiple-choice answer or clicking and dragging.

Evaluation of the psychometric properties of the test

Responses from 233 participants from the third phase of testing using the CT test were analysed to assess the psychometric properties of the test.

Table 39 shows the percentage of students who obtained correct responses to each item of the CT test. This shows that the *Supermarket dash* item was the most difficult question, with only 20% of students getting correct answers. The easiest item was *Select a picture* with 86% correct responses. The correct responses for the items ranged from 20% to 86%.

³⁵ This is referred to as *LIFO ice-cream parlour* rather than '*Ice Cream*' as in the Bebras test item to distinguish it from *ice-cream cones*.

Table 39: Percentage of correct responses for each item on the CT test

Item	% correct
LIFO Ice-cream parlour	29
Park the car	42
Ice-cream cones	70
Watering	72
Select a picture	86
In the forest	52
Supermarket dash	20
Colour machine	53
Magic bracelet	53
Flip-flop	47

Rasch analysis

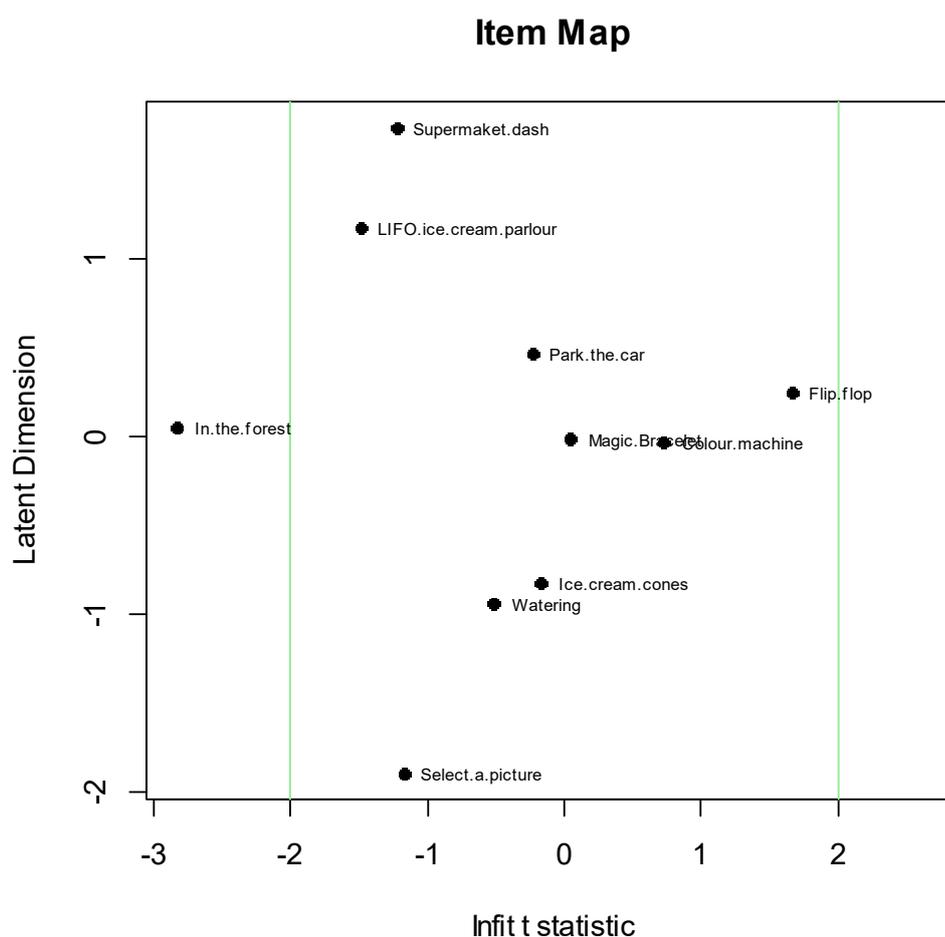
In order to assess the psychometric properties of the CT test, a Rasch analysis was conducted. Rasch analyses represent a method of formally testing aptitude tests and questionnaires against a mathematical model. The mathematical model in question was originally developed by Rasch (1960) and formally relates respondents' aptitudes with item difficulty. Rasch's original model was specifically developed to assess aptitude tests such as the CT test and thus an evaluation against this mathematical model is appropriate for the current data. One of the primary purposes of conducting a Rasch analysis is to assess a test or measure for unidimensionality, that is, whether or not the items are measuring the same underlying aptitude or trait.

The Rasch analysis for the CT test was conducted using the eRm package in the R statistical environment (Mair et al., 2015). This analysis suggested that all items were measuring a single underlying aptitude as evidenced from the non-significant Andersen Likelihood Ratio test ($\chi^2=4.14$, $df=9$, $p=.902$). An additional assessment of unidimensionality is provided by the Martin-Loef test which involves splitting the test into two sub-tests and statistically comparing the two sub-tests. If the items were measuring the same underlying aptitude we would expect the sub-tests to be similar in terms of their fit to the underlying Rasch model. For the Martin-Loef analysis of the CT test using a median-split of the items, there was no significant difference between the two sub-tests in terms of fit to the Rasch model ($\chi^2=25.45$, $df=24$, $p=.381$). The Martin-Loef analysis using a mean and also a random split of the items also produced non-significant differences between the two sub-tests (both $ps>.05$).

An important feature of Rasch analysis is that it can provide an evaluation of the degree to which each item fits the underlying model through item 'infit' and 'outfit' statistics. These fit statistics give an indication of the amount of error between the actual scores from participants for each item and that predicted by the Rasch model. Infit statistics are deemed as most appropriate for assessing how well each item fits the underlying model as these are weighted more heavily towards participants whose performance is closest to the particular item's difficulty level. The outfit statistic is an unweighted measure-of-fit error. All items with the exception of the *In the forest* item, fit the underlying model within generally accepted limits. That is, all had absolute t-values for infit and outfit statistics of less than 2. The *In the forest* item had t-values of -2.49 and -2.83 respectively for outfit and infit statistics. The Bond-Fox pathway map is presented in Figure 12 and this shows each item's infit t-statistics plotted against item difficulty. This clearly shows that all items except the *In the forest* item are well within the accepted limits of infit t-values and are nicely spread across the latent dimension in terms of item difficulty, with the *Select a picture* item being the easiest and the *Supermarket dash* item being the most difficult. The figure also clearly illustrates how the *In the forest* item is quite some way separated from the other items in terms of its infit t-value. This suggests that this item fits the underlying

model a little too well (Bond & Fox, 2015). However, accepted wisdom suggests that overfitting items are less of a threat to validity than under-fitting items. Also, Linacre (2002) suggests that infit and outfit mean square values between 0.5 and 1.5 are 'productive for measurement' and thus this suggests that although *In the forest* has significantly misfitting infit and outfit statistics, the mean square values for this item (0.81 and 0.85 for outfit and infit respectively) are such that we can reasonably leave this item as part of a valid and reliable computational thinking test.

Figure 12: Bond-Fox pathway map for the infit t-statistics



A Rasch analysis can provide a measure of internal reliability akin to the traditional Cronbach's alpha. This is the Person Separation reliability coefficient. For the CT test, the Person Separation reliability was 0.63 which whilst not particularly high, is suggestive of a reasonable level of internal consistency (Wright & Stone, 1999, suggest that this level equates to moderate Person Separation reliability). This value is contrasted with Cronbach's alpha for the CT test. However, as the data for the CT test were dichotomous values (correct or incorrect) the alpha was calculated with the tetrachoric correlation matrix rather than the traditional Pearson correlation matrix (see Gaderman et al., 2012). The Cronbach's alpha for the CT test was 0.72 which again suggests only a reasonable level of internal consistency. Given that the *In the forest* item had a high item fit t-value, the reliability analyses were repeated for the CT test measure without the *In the forest* item being included. The Person Separation reliability coefficient dropped markedly to 0.57 and similarly the Cronbach's alpha dropped to 0.67. It thus appears that the CT test is more internally reliable if the *In the forest* item is retained.

Correlations of computational thinking test scores with Key Stage 2 grades: pilot

The analysis used data from 233 students who took the CT test and for whom we obtained KS2 maths, reading and writing grades. Initial analyses of outliers were undertaken by generating Cook's distance scores

for the correlation between total CT test scores and KS2 maths scores. Plotting Cook's distance scores, as recommended by Cohen et al. (2003), highlighted two students with significantly influential scores. Given the large sample size, it was felt appropriate to remove these students' scores from the analyses. Both the CT test scores and the KS2 maths scores were reasonably normally distributed, and so a Pearson's Product Moment Correlation coefficient was calculated to examine the strength of the relationship between the two measures. This showed that there was a moderate correlation between CT test scores and KS2 mathematics scores ($r=0.45$, $n=231$, $p<.001$). An additional analysis was conducted to establish the relationship between CT test scores and KS2 reading and writing grades. The KS2 writing scores were approximately normally distributed and there were no unduly influential scores. The analysis of these showed that the relationship between CT test and writing scores was lower than for KS2 mathematics but still statistically significant ($r=0.37$, $n=231$, $p<.001$). As the reading scores were heavily negatively skewed, a Spearman's Rho coefficient was calculated and as with the writing scores there was a statistically significant correlation between CT test and reading scores (Rho=0.31, $n=231$, $p>.001$) but this was smaller than the correlation for mathematics.

It is clear from these analyses that although the correlations are moderate, the CT test scores are related to not only mathematical attainment as measured at KS2, but also reading and writing attainment. To establish which of these variables is most important in relation to CT test scores, a multiple regression analysis was conducted with CT test score as the criterion variable and KS2 maths, reading and writing scores as independent variables. This produced a statistically significant regression model ($F(4,227)=20.83$, $p<.001$) with a multiple R of 0.47 and adjusted R^2 of 0.21, suggesting that KS2 mathematics, reading and writing grades together account for 21% of the variation in CT test scores. Interestingly, of the individual KS2 grades, only KS2 mathematics was a significant predictor of CT test scores ($b=0.16$, $t=4.00$, $p<.001$). The regression coefficients for KS2 reading and writing were not statistically significant. This shows that only KS2 mathematics grades have a statistically significant unique relationship with CT test scores. It should be noted that a similar pattern of results was observed when relating CT test scores with KS1 grades but these were a little weaker compared to the KS2 analyses³⁶. This would be expected given the time lag with KS1 tests.

Limitations

It is important to note that the test development has a number of limitations. Perhaps the main limitation is the moderate level of internal reliability. The reliability coefficients highlight that this aspect could be improved. The moderate reliability can be related to the constraints in terms of time and resources for the computational thinking test development. Ideally the initial bank of questions for the test would have been larger to allow for better identification of age-appropriate and test-suitable items. Additionally, we required a test which could be delivered online and which could be completed within 30 minutes. This necessarily limited the type of questions that could be included and the number of questions on the test. Usually internal reliability is directly related to the length of the test in terms of number of questions. Having to limit the length of the test to only 10 items partially explains why we can report here only moderate levels of internal reliability.

A further potential limitation is that the *In the forest* item may not fit too well with the rest of the test items. This was indicated by the high infit t-statistics from the Rasch analysis. However, removing this item leads to considerably lower reliability statistics (e.g. person separation reliability and Cronbach's alpha) than with the item included. Thus, overall it seems that it is better to include rather than exclude this item from the final test.

Although there was some consideration of question ordering when compiling the test, perhaps more attention could have been devoted to this. An alternative approach, and one taken in use of Beaver items in the code club evaluation (Straw, Bamford, and Styles, 2017), was to use random ordering of questions for each participant.

We included what was observed in the initial piloting as one of the easier questions at the beginning of the test to try to ensure that test takers themselves felt that they had the ability to answer the test questions

³⁶ The Pearson correlation of CT test with KS1 maths was 0.36, with KS1 writing 0.32, and with KS1 reading 0.28.

correctly. This was important as we wanted test takers to attempt all questions to the best of their ability. If the initial questions had been too hard then it may have impacted negatively on test takers' engagement with later questions. We could however have considered more carefully the ordering of the later test items to ensure that performance on items was not influenced too much by successful completion of earlier items. However, examination of success rates for the test items does not suggest at all that earlier items influence responses to later ones in the test. Additionally, the infit statistics from the Rasch analysis suggested only one item (*In the forest*) which could be said to be overfitting (being too good a fit to the model). Overfitting items are sometimes the result of them being influenced by responses to other items in the test (see Bond & Fox, 2015). However, if this item was being unduly influenced by earlier items in the test, we would perhaps expect a higher success rate than 52% for the item.

Another potential limitation of the current analyses is that we cannot state from the correlational analyses whether computational thinking as measured by the CT test influenced KS2 mathematics scores or whether it is mathematics ability which influences CT test scores or there are mutual influences. Thus, we need to interpret the correlations with KS2 mathematics with some caution. That said, the moderate correlation with KS2 mathematics results does suggest that the CT test is measuring something related to the KS2 Mathematics tests.

Finally, we were not able to interview CT test takers about their responses to each item. Ideally, during the initial piloting of the test, it would have been useful to get more detailed feedback from test takers as to how they were answering the questions and how well they understood what was being asked of them for each one. Feedback on this was however gained from teachers, and this feedback was used to improve the wording of some of the items for the final version of the test used for correlating with KS2 scores.

Further, only a limited range of computational thinking concepts were implicated in test items. In particular, understanding or application of concepts of loops, logical expressions and data storing (Brennan and Resnick, 2012) were not tested. Other important concepts such as parallelism and conditionals, which are important aspects of ScratchMaths Y5 materials, were tested on a limited number of items. Considering ScratchMaths as a mathematics intervention, it is also important to highlight the concept of a variable, which is an important aspect of Y5 ScratchMaths materials, but not tested in the CT test.

Suitability of the test for use in the ScratchMaths trial

Given the limitations of the test highlighted above, how well suited is it as a measure of computation thinking ability for use in the ScratchMaths trial? The indications from the analysis presented here are positive, notwithstanding the limitations already discussed. According to the Rasch and reliability analyses, all the items appear to be measuring the same underlying construct, and this construct (total score on the test) appears to be moderately related to mathematics ability. This latter correlation adds to the construct validity of the test as we would expect computational thinking skills to be related to mathematics ability. This construct validity is further supported by the regression analysis which suggest that only KS2 mathematical ability rather than reading or writing ability had a significant unique relationship with computational thinking scores. As the CT test total score was related to mathematical thinking, this supports the underlying theory of change for the intervention and supports the utility of the CT test for use in the ScratchMaths trial.

However, the ScratchMaths team raised concerns about the validity of the CT test and its robustness in the context of the national computing curriculum. Specifically, the ScratchMaths team contended that the items generally tested the pre-formal programming stage and were selected from the Beaver tests 2008, 2013 and 2014 before the national computing curriculum was introduced. Thus, it was argued, the items were not well-matched to its requirements and thus what the ScratchMaths intervention endeavoured to foster. Further, there were doubts about whether the items could differentiate between those pupils who had progressed in computational thinking and those who had not, in the intervention and control groups, both of which were following the national curriculum. In addition, there were several key concepts not tested by the CT test that might have strengthened its validity: such as using commands with inputs; using random inputs; using a range of control structures; predicting the outcome of a given formal description of a behaviour/process; parallel scripts for multiple actors.

Some of the ScratchMaths team's concerns reflect limitations in the test discussed above and so reinforce caution about interpreting findings of differences in test outcomes. However, it is reasonable to assume that developing multistructural and relational computational thinking (or, in the ScratchMaths team's terms, formal computational thinking) would entail also the development of the unistructural (pre-formal) level. This, though, has not been formally established. In relation to the national computing curriculum in England, the Beaver/Bebras test items are designed for use internationally and are not specific to any computing curriculum. Analysis of test items and composition for 2015 and 2016 does not suggest any discernible change in the level of challenge of the age-related tests. After the main trial testing the ScratchMaths team suggested that items that are on the junior (age 11-12) Beaver/Bebras test but which do not appear on the test for primary children should have been considered for inclusion, as these tested more formal computational skills. However, given the overall outcome of normal distribution around the middle of the test scale, we believe that if we had done this, the distribution would likely have been skewed and made the test less sensitive across the range of attainment. ScratchMaths may have a positive impact on aspects of computational thinking that are not tested by the CT test and it is important to recognise that it is not a test of programming knowledge or skill. However, given that the test did identify a difference in outcome then this supports the view that it was suitable for use in the trial.

Computational thinking test items included in the final version of the CT measure.

Question 1:

At the LIFO ice-cream parlour the scoops of ice-cream will be stacked in your cone in the exact order in which you ask for them.

What do you have to say in order to get the ice-cream shown in the picture?

I would like an ice-cream with....



Chocolate

Smurf

Strawberry

Drag the flavours into the correct order:

Smurf

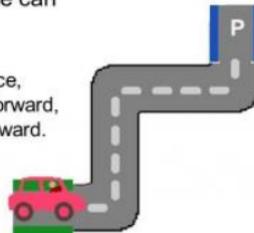
Chocolate

Strawberry

Question 2:

The path of the car from its position to the parking place can be described by the following commands:

- forward** - car will move to the next turn or to the parking place,
- right** - car will turn right by 90 degrees, without moving forward,
- left** - car will turn left by 90 degrees, without moving forward.



Please put these commands in the correct order to take the car to the parking space.

Forward

Forward

Forward

Forward

Left

Left

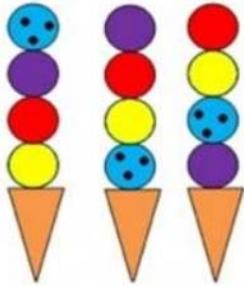
Right

Question 3:

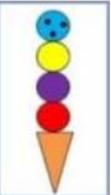
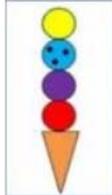
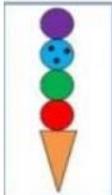
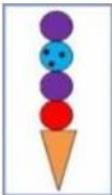
A special ice cream machine fills ice cream cones with four scoops.

It does this in a systematic way.

The picture shows, from left to right, the last three ice cream cones created by the machine.

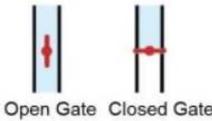


Which of these ice cream cones will be produced next?



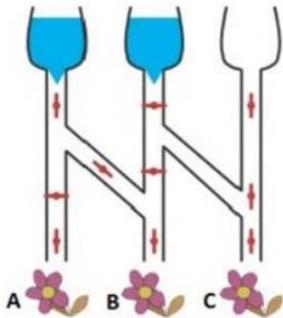
Question 4:

The diagram shows how a watering system is connected. The system consists of tubes and gates. Open and closed gates are shown in the diagram by the direction of the switch.



Water only flows through open gates.

Which of the flowers will get watered when the gates are in the positions below?



Select the flower or flowers which get watered:



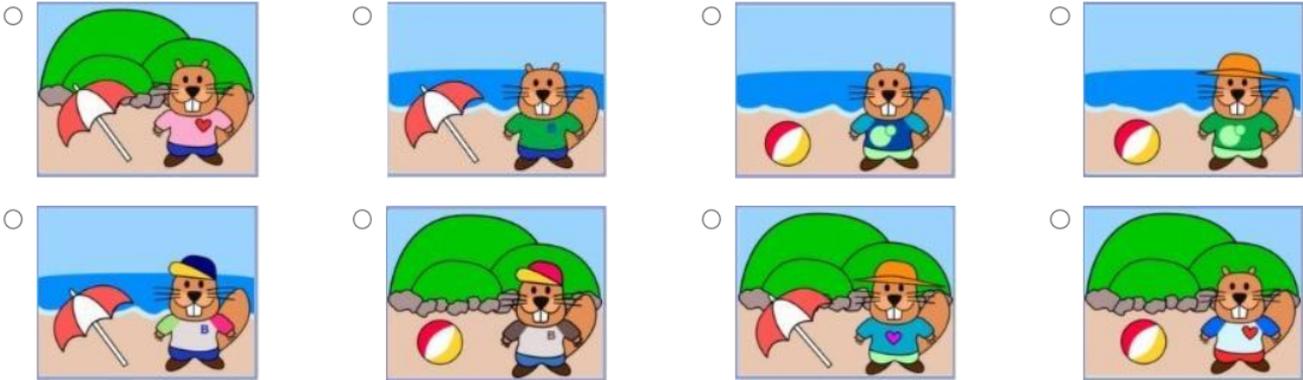
Question 5:

Johnny has 8 photos. He wants to give one to Meera.

He asks Meera three questions to help him select the best picture.

Johnny's Question	Meera's Answer
Do you want a photo with a beach umbrella?	Yes
Do you want a photo where I wear something on my head?	No
Do you want a photo where you can see the sea?	Yes

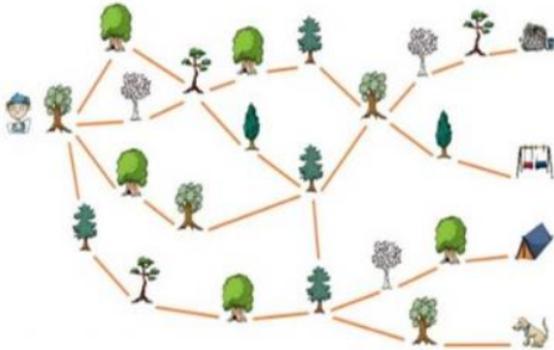
Which photo should Johnny give to Meera?



Question 6:

Rupert went for a walk in this forest.

There were several possible paths but the one Rupert chose led him to a dog.



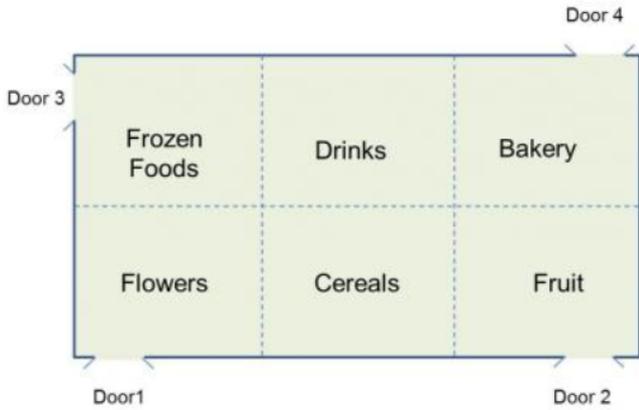
Below are four possible paths Rupert may have taken. Which is the one that led him to the dog?

(The trees are presented in the order in which Rupert saw them.)



Question 7:

At the supermarket I have picked a broken trolley. It only goes straight or turns left. It can't turn right. I need to go to every section of the supermarket



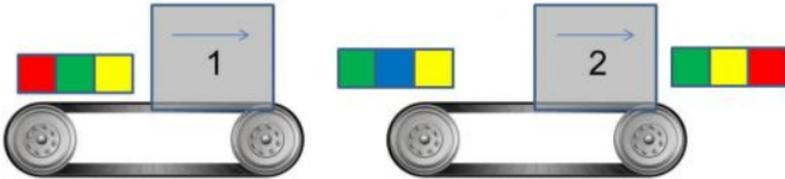
Please select the door I need to enter by and the one I need to leave by..

	Door 1	Door 2	Door 3	Door 4
Enter by:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Exit by:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Question 8:

This machine changes the colours on patterns that are fed into it. The machine has two stages:

In stage 1 green colours get changed to blue and red colours get changed to green.
In stage 2 yellow gets changed to red and blue gets changed to yellow.



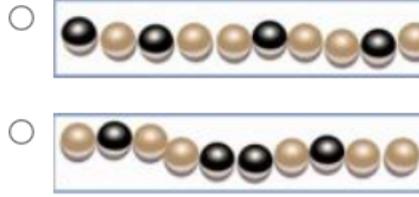
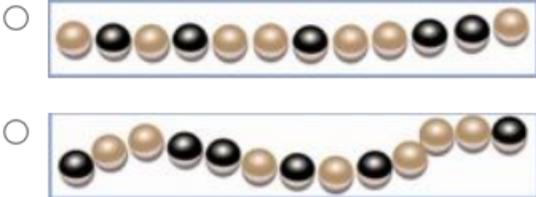
This pattern is fed into the machine:  What comes out?

			
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Question 9:

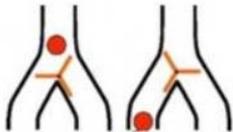


Which of the four bracelets below is the same as the one above?



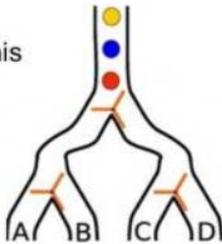
Question 10:

A flipflop is something that is always in one of two states. Every time a ball comes through a flipflop the state changes. The flipflops here work as follows:



The ball falls from the top and then goes either left or right depending on the state of the flipflop. While falling, it pushes the flipflop so that the next ball falls in the other direction

Suppose you had a machine with This arrangement of flipflops:



Where will the yellow ball come out?

- A B C D

Appendix G: Distribution of primary outcome (overall K2 maths attainment, 2017) and follow-on secondary outcomes (attainment in the three KS2 maths papers, 2017)

The primary outcome was obtained from the National Pupil Database (NPD) in December 2017. The specific variable specified in the Statistical Analysis Plan (SAP) for the ScratchMaths trial was 'KS2_MATMRK'. This was the raw total score obtained from summing the scores of three KS2 maths tests: Paper 1 (arithmetic); Paper 2 (reasoning 1) and Paper 3 (reasoning 2). In addition to the specified raw total, the NPD provided a scaled version of KS2 maths attainment ('KS2_MATSCORE'). This scaling is done to 'ensure ... accurate comparisons of performance over time' can be made and that a score of 100 indicates when a pupil has met expected standards.³⁷

Figure 13 summarises the distribution of the raw KS2_MATMRK and scaled KS2_MATSCORE variables.

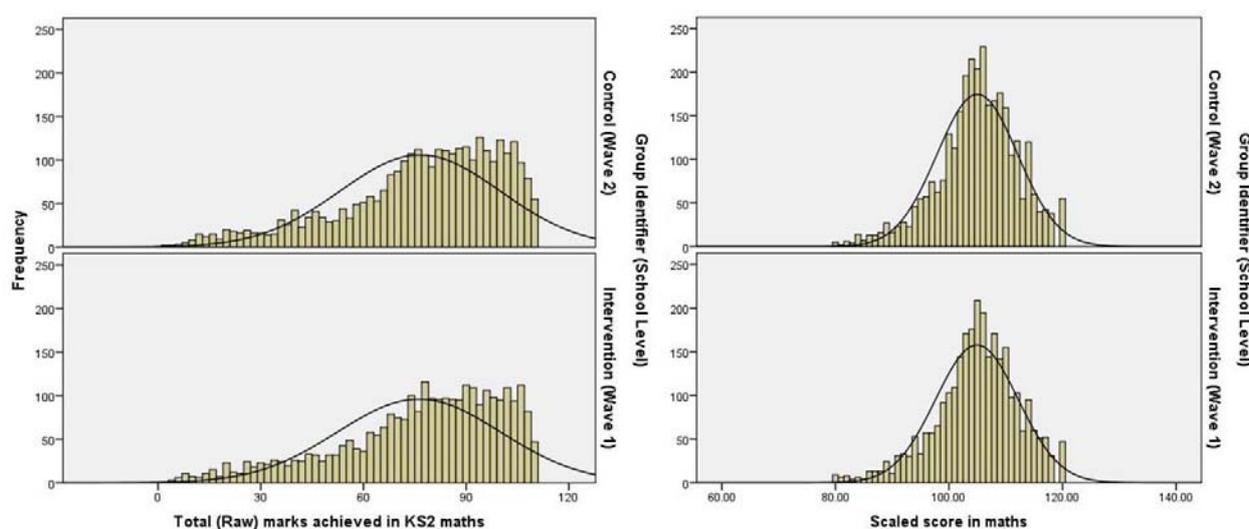
Figure 13: Summary of the distribution of KS2 maths attainment

Raw KS2_MATMRK

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	133 (4.1%)	3,111	76.5 (23.46)	81	0 :110
Intervention	105 (3.5%)	2,877	76.2 (23.85)	80	3 :110
All	238 (3.8%)	5,988	76.4 (23.64)	81	0 :110

Scaled KS2_MATSCORE

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	136 (4.2%)	3,108	105.0 (7.09)	105	80 :120
Intervention	105 (3.5%)	2,877	104.9 (7.26)	105	80 :120
All	241 (3.8%)	5,985	105.0 (7.17)	105	80 :120



³⁷ See <https://www.gov.uk/guidance/scaled-scores-at-key-stage-2> for details on this scaled KS2 maths outcome

As can be seen from the histogram, the raw score distribution has a notable negative skew whilst the scaled score follows a Gaussian distribution more closely. This skew has the potential to bring problems within the multilevel regression analyses. Whilst regression modelling does not assume that the dependent variable follows a normal/Gaussian distribution, there is an assumption that the residuals from the modelling will do so.

Our response to finding this was to maintain the KS2_MATMRK raw score as the primary outcome but also to run parallel analyses using the scaled KS2_MATSCORE outcome. In all cases, we found that the models using either the KS2_MATMRK raw score or scaled KS2_MATSCORE outcome were in agreement and so no additional details are provided.

As noted earlier, the KS2_MATMRK raw score is derived by summing the scores on three KS2 maths assessments. These three assessments are used as secondary outcomes within the ScratchMaths impact evaluation and their distributions are summarised in Figure 14 below.

Figure 14 shows clear negative skews in the distributions for KS2 maths Paper 1 (arithmetic) and Paper 2 (reasoning 1) but this is less evident for Paper 3 (reasoning 2). In addition to the negative skew, ceiling effects are also evident in the distributions for all three KS2 maths papers - most strikingly for Paper 1 which shows a steep 'ski-slope' shape.

The skews and ceiling effects present in the distributions for KS2 maths Papers 1 to 3 present similar problems for regression modelling as the raw primary outcome. We present the distributional details to aid the critical interpretation of the impact evaluation findings.

However, we found no evidence of a statistically significant impact for ScratchMaths on KS2 mathematics attainment overall (using both raw and scaled versions of the outcome) and this (no impact) finding was also found within each of the three KS2 maths assessments.

Figure 14: Summary of the distribution of the three KS2 maths assessments

Paper 1 (Arithmetic)

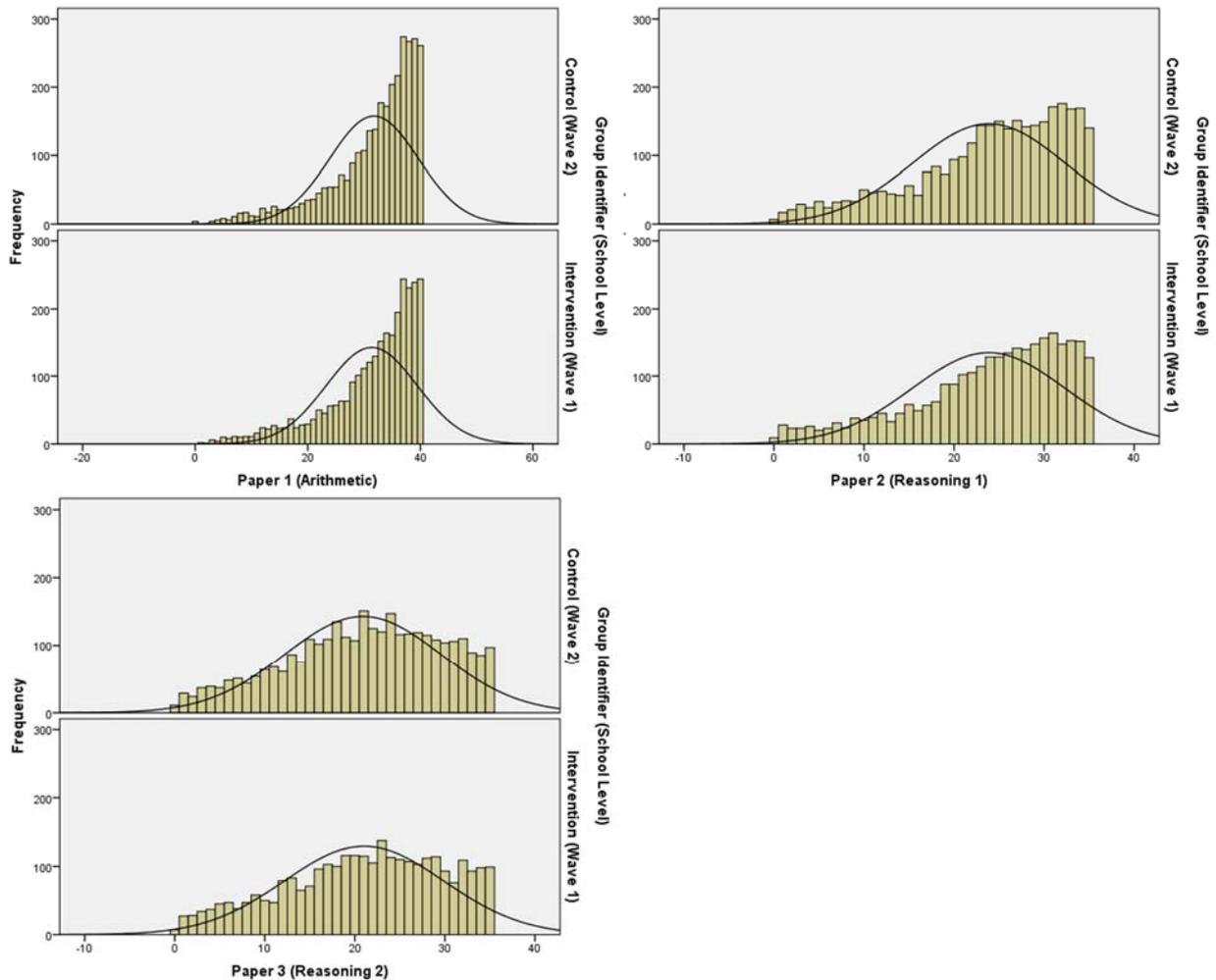
	missing (%)	n=	mean (sd)	Median	Min : Max
Control	132 (4.1%)	3,112	31.8 (7.87)	34.0	0 : 40
Intervention	105 (3.5%)	2,877	31.4 (8.08)	34.0	1 : 40
All	237 (3.8%)	5,989	31.6 (7.97)	34.0	0 : 40

Paper 2 (Reasoning 1)

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	132 (4.1%)	3,112	23.9 (8.48)	25.0	0 : 35
Intervention	104 (3.4%)	2,878	23.8 (8.54)	26.0	0 : 35
All	236 (3.7%)	5,990	23.9 (8.51)	26.0	0 : 35

Paper 3 (Reasoning 2)

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	133 (4.1%)	3,111	20.9 (8.68)	22.0	0 : 35
Intervention	103 (3.4%)	2,879	21.0 (8.77)	22.0	0 : 35
All	236 (3.7%)	5,990	21.0 (8.77)	22.0	0 : 35



Appendix H: Distribution of interim secondary outcome (Computational thinking test, 2016)

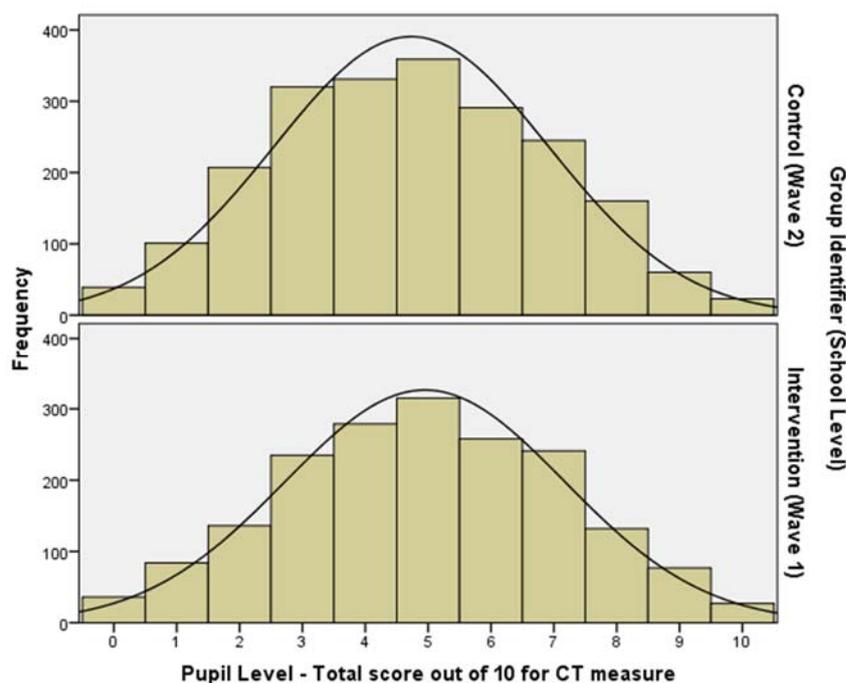
The computational thinking test (CT test) was developed by us and participant pupils took the CT test in June 2016. See Appendix F for more details on the how the CT test was developed. This Appendix H provides a summary of the distribution of CT test scores and the 10 questions from which the overall CT test score was derived.

Figure 15 summarises the distribution of CT test scores. Whilst the CT test scores technically are discrete data (taking integer values between 0 and 10), the distribution follows a normal/Gaussian distribution quite closely.

Figure 15: Summary of the distribution of CT test scores

CT test score

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	1,108 (34.2%)	2,136	4.73 (2.18)	5.0	0 : 10
Intervention	1,162 (38.9%)	1,820	4.95 (2.22)	5.0	0 : 10
All	2,270 (36.4%)	3,956	4.83 (2.20)	5.0	0 : 10



The CT test had a sizable issue with missing data. Much of this relates to whole schools dropping out of the intervention or otherwise not administering the CT test³⁸. No CT test data were obtained from 15 intervention schools (1,162 pupils) and 14 control schools (1,108 pupils). In all, we had CT test data for 40 intervention schools (1,820 pupils) and 41 control schools (2,136 pupils). Among the 81 schools where the CT test took place, the pupil-level response was very good for both the intervention (85% response) and the control (88%) groups.

³⁸ For reasons for school drop out of the intervention see page 58. Reasons given for not completing the CT test were similar - such as changes of staffing and changes of school priorities

The initial multilevel impact analyses used the raw CT test score summarised in Figure 13. The propensity-score-paired-school-stratification research design enabled a follow-on sensitivity analysis using a restricted sample of schools.

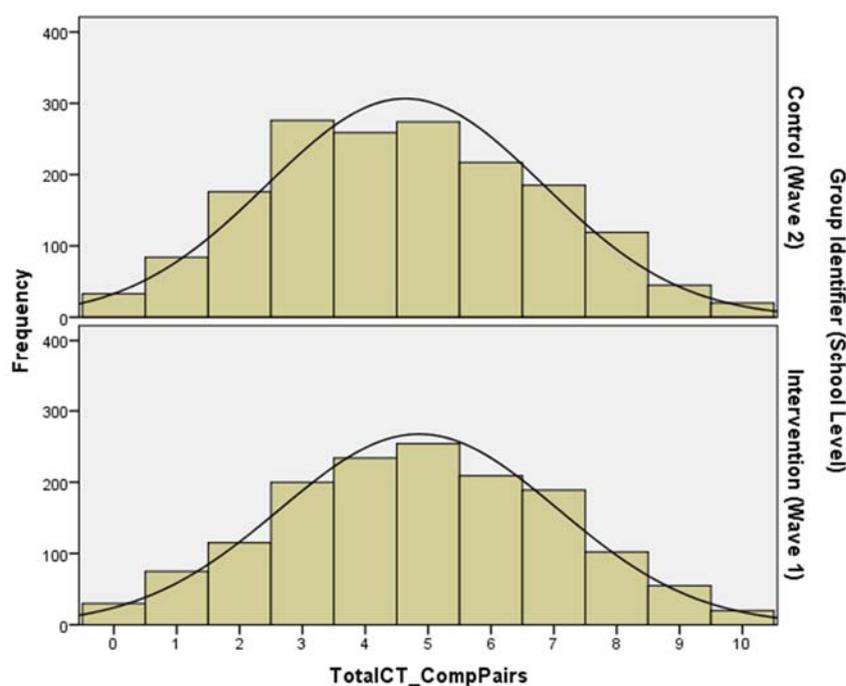
The sample was restricted to include only cases from 'complete pairs' of schools. Specifically, the restricted sample included data only when it was available from both intervention *and* control schools across the initial 55 school pairings. Within the 'raw sample' model reported in Figure 13 above, cases are drawn from 40 intervention and 41 control schools (81 of the 110 schools in the study). However, when taking school pairings into account, within this raw sample of 81 schools, there are nine instances of CT test data being present from intervention schools but not for their matched control school pairs. Further, there are 10 instances of CT test data being present from control schools but not for their matched intervention school pairs. By removing these 19 schools, the sample is restricted to 31 school 'complete pairs' (62 schools in total). The purpose of doing this is to best ensure a good baseline balance between the intervention and control school samples. Schools were paired together prior to randomisation, and so restricting the analyses to just 'complete pairs' should ensure the best baseline balance whilst maintaining the integrity of the RCT design (albeit with a reduction in sample size and hence statistical power).

Figure 16 below summarises the distribution of CT test scores for the restricted 'complete pairs' sample of 62 schools. This shows a similar normal/Gaussian distribution to that seen with the raw CT test score in Figure 13.

Figure 16: Summary of the distribution of CT test scores. Restricted 'complete-pairs' sample of 62 schools

CT test score

	missing (%)	n=	mean (sd)	Median	Min : Max
Control	1,483 (46.8%)	1,688	4.64 (2.20)	5.0	0 : 10
Intervention	1,503 (49.1%)	1,483	4.85 (2.21)	5.0	0 : 10
All	2,986 (47.9%)	3,171	4.74 (2.21)	5.0	0 : 10



Appendix I: Multilevel analyses & calculation of effect sizes

This Appendix provides additional model details for the intention-to-treat (ITT) impact analysis of ScratchMaths on KS2 maths attainment. The Appendix also explains how the model coefficient was converted into the effect size statistics shown in the main report. Further model details are also provided for the impact analysis of ScratchMaths on the interim CT test outcome. At the end of the Appendix is an example of the STATA code that was used for these analyses. Fuller model and STATA Do-files can be provided on request.

Headline ITT analysis of KS2 maths attainment

Table 40 summarises the multilevel ITT impact analyses for the KS2 maths primary outcome.

Table 40: Main Impact Analyses Models for Primary Outcome - KS2 Maths Attainment

Description	Stage 0 Empty Model		Stage 1 Outcome Only		Stage 2 Impact Model		
	coef.	s.e.	coef.	s.e.	coef.	s.e.	p=
Group (Intervention)	-	-	-0.25	1.644	0.06	1.471	0.970
Pre-test (KS1 maths) at pupil level	-	-	-	-	5.02*	0.066	<0.001
Pre-test (KS1 maths) aggregated to school level	-	-	-	-	-1.02	0.661	0.122
Constant	76.04	0.822	76.16	1.158	75.54	1.036	
Variance decomposition (s.e.)							
<i>School Level</i>	61.6 (10.67)		61.6 (10.66)		50.5 (8.36)		
<i>Class Level</i>	4.4 (3.78)		4.4 (3.78)		6.2 (2.55)		
<i>Pupil Level</i>	500.1 (9.30)		500.1 (9.30)		244.9 (9.30)		
<i>Total</i>	566.1		566.1		301.6		
ICC Statistics							
<i>School Level</i>	0.11		0.11		0.17		
<i>Class Level</i>	0.01		0.01		0.02		
Explanatory Power							
<i>School Level</i>	-		0.00		0.18		
<i>Class Level</i>	-		0.00		-0.40		
<i>Pupil Level</i>	-		0.00		0.51		
<i>Total</i>	-		0.00		0.47		
Sample Sizes							
<i>Number of Schools</i>	110		110		110		
<i>Number of classes</i>	207		207		207		
<i>Number of Pupils</i>	5,988		5,988		5,818		
Effect Size (Hedges g)							
	-		-0.01		0.00		
95% CIs:							
<i>Lower</i>	-		-0.15		-0.12		
<i>Upper</i>	-		+0.13		+0.12		

* p<0.05

Calculating Hedges g effect size

As specified in the statistical analysis plan, the impact of ScratchMaths was measured using the Hedges g effect size statistic based on the formula shown below.

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta_s^2 + \delta_c^2 + \delta_p^2}}$$

Where δ_s^2 is the school level variance, δ_c^2 is the class level variance and δ_p^2 is the pupil level variance for the (stage 0) empty model and $(T - C)_{adjusted}$ is the coefficient estimate for the group identifier dummy variable from the (stage 2, impact) model.

From Table 40, the total variance ($\delta_s^2 + \delta_c^2 + \delta_p^2$) is 566.1 and so the standard deviation $\sqrt{\delta_s^2 + \delta_c^2 + \delta_p^2}$ is 23.27.

Also from Table 40, $(T - C)_{adjusted}$ is 0.06.

$$ES = \frac{(T - C)_{adjusted}}{\sqrt{\delta_s^2 + \delta_c^2 + \delta_p^2}} = \frac{0.06}{23.27} = 0.0026 \sim 0.00 \text{ standard deviations}$$

The upper and lower 95% confidence intervals for the coefficient are similarly divided by 23.27 to convert them into confidence intervals for the effect size.

Example of STATA Code used in multilevel analyses

Stage 0: `mixed KS2_MATMRK if (OptOuts==0) || LEAESTAB: || CLASS_IDEst:`

Stage 1: `mixed KS2_MATMRK GroupXX if (OptOuts==0) || LEAESTAB: || CLASS_IDEst:`

Stage 2: `mixed KS2_MATMRK GroupXX KS1MATPOINTS_CENT KS1Maths_SchCENT if (OptOuts==0) || LEAESTAB: || CLASS_IDEst:`

As discussed in the main report, an additional sensitivity stage was also undertaken:

Stage 3: `mixed KS2_MATMRK GroupXX KS1MATPOINTS_CENT KS1Maths_SchCENT
KS1AttainSCHOOL MathsKS1to2VASCHOOL TOTPUPSSCHOOL
GENDER_GIRLSSCHOOL FSMSCHOOL EALSCHOOL SENSCHOOL
b2.pairhub_AUT if (OptOuts==0) || LEAESTAB: || CLASS_IDEst:`

The STATA Do-file will be archived with the data and is available on request.

Analysis of interim CT test outcome

Table 41 summarises the multilevel impact analyses for the interim CT test outcome.

Table 41: Main impact analyses models for interim CT test outcome

Description	Stage 0 Empty Model		Stage 1 Outcome Only		Stage 2 Impact Model		
	coef.	s.e.	coef.	s.e.	coef.	s.e.	p=
Group (Intervention)	-	-	0.24	0.195	0.33*	0.165	0.048
Pre-test (KS1 maths) at pupil level	-	-	-	-	0.30*	0.009	<0.001
Pre-test (KS1 maths) aggregated to school level	-	-	-	-	0.15	0.077	0.059
Constant	4.86	0.098	4.74	0.136	4.78	0.114	
Variance decomposition (s.e.)							
<i>School Level</i>	0.64	(0.125)	0.62	(0.123)	0.41	(0.087)	
<i>Class Level</i>	0.10	(0.048)	0.10	(0.048)	0.12	(0.044)	
<i>Pupil Level</i>	4.15	(0.095)	4.15	(0.095)	3.15	(0.074)	
<i>Total</i>	4.88		4.87		3.70		
ICC Statistics							
<i>School Level</i>	0.13		0.12		0.11		
<i>Class Level</i>	0.02		0.05		0.03		
Explanatory Power							
<i>School Level</i>	-		0.02		0.36		
<i>Class Level</i>	-		-0.01		-0.17		
<i>Pupil Level</i>	-		0.00		0.24		
<i>Total</i>	-		0.003		0.25		
Sample Sizes							
<i>Number of Schools</i>	81		81		81		
<i>Number of classes</i>	162		162		162		
<i>Number of Pupils</i>	3,956		3,956		3,841		
Effect Size (Hedges g)							
	-		+0.11		+0.15		
95% CIs:							
<i>Lower</i>	-		-0.06		+0.001		
<i>Upper</i>	-		+0.28		+0.29		

* p<0.05

The effect size of +0.15 was calculated by dividing the 0.33 Group (Intervention) model coefficient by the square root of 4.88 (total variance for empty model).

Appendix J: Process evaluation samples and data consolidation

Process evaluation samples and compositing

Attendance data

Two sets of data were collected for ScratchMaths professional development attendance: from IoE registers and from the two implementation and process evaluation (IPE) surveys. The IoE data for this dimension of fidelity are preferred as more reliable because they were collected at the time of attendance. It is also more complete.

Table 42 compares the survey and IoE attendance data. This is presented, firstly, for the composited sample of respondents to the Y5/Y6 surveys (n=36 in Y5, n=31 in Y6). As discussed in the section on methods, composited responses refers to the process by which, in some cases, survey responses by more than one teacher in a school was reduced to a single school record. The approach to compositing is described in detail in Appendix K. Secondly, the sample is restricted so that only schools where we have data for both Y5 and Y6 are included (n=27).

For this fidelity dimension, when attendance data from two teachers were present, the mean attendance of the two was used.

Table 42: Comparing ScratchMaths team and composited survey responses on attendance data at the school level

	Y5		Y6	
	IoE attendance record	Y5 Survey responses	IoE	Y6 Survey responses
n=	36	36	31	31
Mean days (sd)	2.2 (1.03)	2.2 (0.75)	2.6 (1.42)	1.8 (0.92)
IoE>Survey	16 (44%)		16 (52%)	
IoE=Survey	9 (25%)		10 (32%)	
IoE<Survey	11 (31%)		5 (16%)	

In Y5, whilst the mean attendance is the same, only 25% of the data shows agreement between the IoE records and the survey, and in Y6 the mean attendance recorded by IoE is greater for the 31 schools than that recorded in the survey. For Y6 there is only agreement in 32% of cases. This is likely to be due to issues of not recalling accurately participation in PD. However, it does indicate that responses to other survey questions should be treated with caution.

Survey responses

Table 43 below provides details of survey returns. The 'Responses' column shows the total number of attempts to complete the survey regardless of whether some of these were by the same person.. In three of the five surveys there was more than one response by a single teacher or the number of survey items completed was not sufficient to include the data in analysis. Thus, the 'Teachers' column represents the number of analysed responses. In some cases, there was also more than one response per school, detail on this is given in Appendix K.

Table 43: Survey samples

Survey	Timing	Purposes	Responses	Teachers	Schools
Wave 1 Y5	Summer 2016	Implementation and fidelity data	48	44	36
Wave 1 Y6	Summer 2017	Implementation and fidelity data	35	33	31
Wave 2 Y5 (control)	Summer 2016	Assess control condition	40	37	34
Wave 2 Y6 (control)	Summer 2017	Assess control condition	1	1	1
Wave 2 Y5 (waitlist)	Summer 2017	Data to inform RQ7, RQ8, RQ9 and scalability	13	13	13

As shown, there was only one completion of the Wave 2 Y6 control survey. The ScratchMaths team did not have contact details for the Wave 2 Y6 teachers and so they were contacted via the Y5 Wave 2 teachers.

Table 44 provides details of the relationship between interview participants and their completion of the Wave 1 and Y5 surveys for Wave 1 (intervention schools).

Table 44: Interview participants their survey responses

Interview year	Number of schools	Wave 1 intervention survey response (Y5)	Wave 1 intervention survey response (Y6)
2016	7	7	2
2017	7	7	7
2016 & 17	2	2	2

All schools that provided interviewees in 2017 had a teacher complete both the Y5 survey and then the Y6 survey. Two of the schools who provided interviewees in 2016 also provided an interviewee in 2017. Five schools which were interviewed in 2016 did not complete the survey in 2017. There were 39 schools for which there is Wave 1 survey data in either Y5 or Y6 or in both years.

Using multiple responses and compositing surveys

In cases where a teacher had responded more than once to a single survey, responses were composited into a single teacher record. The term 'composited' is used as a general term for the process of combining different though related items.; if there were any differences between responses to a single item by the same teacher then the last data provided were preferred (in most cases multiple attempts were due to first responses not being full attempts to complete the survey). There are instances where we had two or more teacher responses within a single school; in Y5 there were eight cases of this and in Y6 there were three cases. Details on how multiple responses from a single school were composited are reported in Appendix K.

Bias in survey and interview samples

The 16 schools that were classified as not having sustained participation did not participate in surveys or interviews apart from a single school that had one teacher who completed the Wave 1 Y6 survey. This means that there is a likely bias in the survey responses, with those schools that participated less, and so, potentially, having less favourable views of ScratchMaths professional development and materials being under-represented.

Appendix K: Fidelity to ScratchMaths & the on-treatment analysis

Determining fidelity

Detail of fidelity criteria and how they were developed are included in the implementation and process evaluation methods section..

Fidelity was measured across both years of years of the trial and was based on a subsample of 27 of the 55 intervention schools with survey response(s) from teachers in both Y5 and Y6. Of these 27 schools a response for a single teacher was obtained for 21 schools in Y5 and 24 schools in Y6. We received two teacher responses for six schools in Y5 and for three schools in Y6. We did not receive more than two teacher responses per school.

- Teacher level fidelity to the ScratchMaths intervention was specified across both Y5 and Y6 using five dimensions
 - PD Attendance (IoE attendance data)
 - School IT provision (Y5 teacher survey)
 - Use of ScratchMaths Materials (Y5 & Y6 teacher surveys)
 - Curriculum time (Y5 & Y6 teacher survey)
 - Following order of modules (Y5 & Y6 teacher survey)
- Data was collected for all five dimensions from 27 of the 55 ScratchMaths intervention schools (i.e. a response for both the Y5 and Y6 teacher surveys). This teacher level data was used to measure fidelity at the school level in order to bring fidelity into the impact analyses).
- The following criteria were adopted for using the teacher level fidelity data to estimate fidelity at the school level:
 - For 21 of the 27 ScratchMaths intervention schools there was a single teacher response in both Y5 and Y6. In these instances, school level fidelity was constructed using this data (that is the teacher is assumed to represent the school).
 - In six intervention schools there were two teacher responses for the Y5 or Y6 surveys. Whilst taking an average of the two responses was possible for fidelity dimensions defined as a scale measure (PD attendance and use of materials), this was not possible for dimensions defined as a nominal or ordinal measure.

Table 45 summarises the approach taken for compositing the five ScratchMaths fidelity dimensions where there were more than one teacher.

Table 45: Approach to compositing multiple teacher fidelity data to determine a school level

Fidelity component	Approach
PD attendance	Mean attendance for two teachers in Y5 & Y6
IT Provision	IT provision was universally high fidelity across all teacher responses (i.e. all reported at least 2:1 pupil: teacher ratio).
Use of ScratchMaths Materials	Mean coverage of modules for two teachers in Y5 & Y6
Curriculum time	<p>Teachers were asked to identify how much time they spent using ScratchMaths in the classroom and provided with three responses (<12; 12 to <20, 20+).</p> <p>When there was disagreement here (in 4 of the 6 schools this was the case), the teacher that reported a higher number of hours was used to represent the school.</p>
Order of modules	<p>Teachers were asked to respond yes/no to whether they followed the order of ScratchMaths modules.</p> <p>When there was disagreement here (in 1 of the 6 schools this was the case), the teacher that reported 'yes' was used to represent the school.</p>

Limitations

Fidelity criteria were determined by the ScratchMaths team towards the end of the trial, and at the teacher level (see Table 6 in methods section of report). This teacher level data was used to create a school level fidelity measure in a post-hoc way at the analysis stage. Assuming that the response of one teacher represents the school is a limitation of fidelity assessment.

Taking the average (PD attendance, use of materials) from two teachers within a school does mean that the school level measure draws on more information than a single response but also means that for six of the 27 schools, the measurement of fidelity for these two dimensions was 'different' to the other 21.

For the other two fidelity dimensions, when teacher responses disagreed, the response was selected that indicated a higher level of fidelity. This was preferred to the opposite approach - selecting the response indicating a lower level of fidelity. This was because having an indication of high fidelity within a school from at least one teacher was more reflective of the approach taken for the 21 instances when we had a single teacher response.

Data from teacher interviews suggest that where there was variance between classes in schools this was relatively minor, for example, different amounts of time spent on activities or different ways activities were introduced. Thus, it is reasonable to assume that the differences were not substantial and given the findings of the impact and on treatment analysis they are not consequential to the reported findings.

Professional development attendance fidelity

Table 46 shows that 19 of the 27 schools (70.4%), for which there was sufficient data to include in the on-treatment analysis, are identified as having high fidelity for attendance whilst 23 (85.2%) are identified as having high or medium fidelity.

Table 46: School-level PD Attendance in Y5 and Y6

Y5 Fidelity (PD Attendance)	Y6 Fidelity (PD Attendance)		
	Low (<1)	Med (1-2)	High (2+)
Low (<1)	2	0	0
Med (1-2)	2	0	4
High (2+)	0	0	19

Technology use fidelity

Across all schools in Y5, 35 out of 36 met the high criterion and in Y6, 30 out of 31 did so. In the sample of 27 there was a minimum of 2:1 pupil-to-computer ratio - therefore all 27 (100%) reached the threshold of high fidelity for this dimension.

Material use fidelity

The fidelity criteria for use of material are given in Table 48 below.

Table 47 summarises the coverage of the three Y5 and three Y6 ScratchMaths module materials from responses to the implementation and process evaluation teacher surveys. For each module, a mean score is calculated based on responses to the four investigations in each module³⁹; the module test has not been included in these calculations. When more than one teacher provided data, the mean score was taken to represent the school.

Table 47: Mean coverage of ScratchMaths module materials

Restricted	M1 (Tiles)	M2 (Beetle)	M3 (Sprites)	M4 (Build N)	M5 (Explore)	M6 (Geom)
n=	27	27	27	27	27	27
Mean (sd)	0.92 (0.227)	0.87 (0.261)	0.45 (0.405)	0.72 (0.406)	0.50 (0.422)	0.30 (0.386)

Across all six ScratchMaths modules

Overall	Y5	Y6	Y5 & Y6 combined
n=	27	27	27
Mean (sd)	0.75 (0.233)	0.51 (0.347)	0.63 (0.222)

Within the 27 schools with Y5 and Y6 IPE survey data, the overall coverage of ScratchMaths materials from the six modules was 63%⁴⁰. Coverage was higher in Y5 (75%) compared with Y6 (51%). In each year, coverage was highest with the first module of the year but dropped with the subsequent two modules.

³⁹ Within a particular module, if a teacher reported to use two of the four investigations, the mean coverage score was 2/4 (0.5); three of the four investigations = 0.75 etc.

⁴⁰ Each of the six Y5 and Y6 ScratchMaths modules had four investigations (24 investigations in all), an overall coverage of 63% across Y5 & Y6, which means that teachers reported to use around 15 of the 24 (63%) investigations across the six modules.

This dimension of fidelity does not require the fine-grained detail shown in Table 47; the focus is on whether at least some of the core activities were undertaken. Table 48 below summarises this for Y5 and Y6.

Table 48 shows 17 schools (63.0%) with high fidelity (reporting to cover five or all six ScratchMaths modules over the two years) and 19 schools (70.4%) identified as having medium fidelity or higher (reporting to cover four or more ScratchMaths modules over the two years).

Table 48: ScratchMaths module coverage (using the 'at least some' fidelity criterion)

Y5 through Y6	Restricted Sample (n=27 schools)	Coverage Fidelity
No modules covered	0	Low (8)
Just 1 module	0	
2 modules	4 (15%)	
3	4 (15%)	
4	2 (7%)	Medium (2)
5	9 (33%)	High (17)
All 6 modules	8 (30%)	

Time spent teaching ScratchMaths fidelity

Responses for Y5 and Y6 are summarised in Table 49 below, identifying 10 schools (37.0%) with high fidelity and 19 schools (70.4%) with medium or high fidelity.

Table 49: Time spent teaching ScratchMaths in Y5 and Y6

Y5	Y6			
	Less than 12	12 - 20	20+	Missing
Less than 12	0	1	0	0
12-20	1	6	3	2
20+	2	8	2	2

Progression fidelity

Table 50 identifies 22 schools (81.5%) who reported to follow the specified module order in both Y5 and Y6 (i.e. high fidelity).

Table 50: Whether followed order of ScratchMaths modules in Y5 and Y6

	Y6		
Y5	No	Yes	Missing
No	0	1	0
Yes	3	22	1

Overall fidelity

Once the five fidelity dimensions are drawn together:

- Five intervention schools are identified as having high fidelity to the ScratchMaths intervention across all five dimensions in Y5 and Y6.
- 13 schools are identified as having medium or high fidelity.

Table 51 summarises KS2 maths attainment for pupils located in intervention schools identified as having high or medium fidelity to the ScratchMaths intervention during the two-year trial period. Attainment statistics are shown across the separate fidelity dimensions and for the overall high or medium/high fidelity thresholds.

As noted earlier, the fidelity data draw on 27 of the 55 intervention schools. Table 51 begins by comparing attainment of pupils within all 55 intervention schools with the subsample of pupils in the 27 schools with Y5 and Y6 fidelity data. From this, it can be seen that pupils in the 27-school fidelity sample were lower attaining in KS2 maths compared with the complete sample.

From comparing means across different levels of fidelity, there is little evidence of a positive association between fidelity to ScratchMaths and KS2 maths attainment.

Table 51: KS2 maths attainment for pupils in intervention schools identified as having high or medium fidelity]

	KS2 Maths	
	n_p pupils	mean (sd)
<i>ALL Pupils in intervention (wave 1) schools</i>		
Full sample; $n_s=55$ schools	2,877	76.2 (23.85)
Fidelity Sample; $n_s=27$ schools	1,368	74.8 (24.11)
<i>Fidelity Dimension 1 - Y5 & Y6 Attendance</i>		
High/medium ($n_s =23$)	1,145	73.5 (23.97)
Low / other ($n_s=4$)	223	81.9 (23.61)
<i>Fidelity Dimension 3 - Y5 & Y6 Module Coverage</i>		
High ($n_s=17$ schools)	832	73.4 (24.92)
High/medium ($n_s =19$)	944	73.0 (24.82)
Low / other ($n_s=8$)	424	79.0 (21.90)
<i>Fidelity Dimension 4 - Time spent teaching ScratchMaths in Y5 & Y6 Module</i>		
High ($n_s=10$ schools)	497	76.0 (23.06)
High/medium ($n_s =19$)	979	73.7 (24.46)

Low / other (n_s=8)	389	77.7 (22.98)
<i>Fidelity Dimension 5 - Order / Progression of ScratchMaths</i>		
High/medium (n_s =22)	1,108	74.1 (24.43)
Low / other (n_s=5)	260	78.0 (22.47)
<u><i>Overall Fidelity across the five dimensions in Y5 and Y6</i></u>		
High on all dimensions (n_s=5 schools)	277	77.8 (22.92)
High or Medium on all dimensions (n_s=13 schools)	664	71.8 (24.44)
< Medium on all dimensions (n_s=14 schools)	704	77.7 (23.45)

On-Treatment Analysis

The main intention-to-treat (ITT) analyses found no evidence to suggest that ScratchMaths had a positive impact on KS2 maths attainment. The ITT approach included pupils in all intervention schools regardless of fidelity to ScratchMaths. Follow-on on-treatment analyses limit the analyses to the five schools identified as having high fidelity and 13 schools identified as having medium/high fidelity.

The fidelity analyses summarised in Table 51 above do not provide clear evidence that higher levels of fidelity to ScratchMaths are associated with higher levels of KS2 maths attainment. However, follow-on multilevel analyses were undertaken in order to explore impact more comprehensively given that this was an efficacy trial.

Three model stages were used

- Stage 1: ScratchMaths dummy variable, KS1 maths (at pupil and school levels).
- Stage 2: include CT test score.
- Stage 3: include all of the school-level variables that were included as explanatory variables to generate the propensity scores used for randomisation and dummy variables to identify school pairs within geographical hubs.

For these on-treatment analyses, the sample of intervention schools was reduced to just the sample of five high-fidelity schools and then to just the sample of 13 medium/high-fidelity schools. In each of these on-treatment analyses, two comparison groups were used. First, the raw sample of 55 control schools. The second comparison group drew on the propensity-score-paired-school-stratification research design to limit the comparison group of schools to those matched to the five high-fidelity and 13 medium/high-fidelity intervention schools.

Coefficients and effect size estimates from the Stage 1 and 2 models are summarised in Table 52. In most instances a negative effect size is observed but none are statistically significant.

In summary, from these on-treatment analyses we found no evidence to suggest that the ScratchMaths intervention had an impact on KS2 maths attainment. This was the case when just KS1 maths attainment was controlled for and when both KS1 maths attainment and CT test score were controlled for. We therefore conclude that, among the sample of 27 schools where we have IPE teacher survey data in both Y5 and Y6, we found no relationship between fidelity to ScratchMaths and KS2 maths attainment directly (taking account of KS1 attainment prior to randomisation) or indirectly through computational thinking (as measured by the interim CT test in 2016).

Table 52: Summary of on-treatment models used to evaluate the impact of ScratchMaths on KS2 maths attainment**High-fidelity subsample (five intervention schools)**

Coefficient & Effect Size from:						
<ul style="list-style-type: none"> stage 1 (KS1 maths) stage 2 (KS1 maths & CT test) 						
Sample	Stage	n in model (intervention; control)	Coef (95% CI)	Hedges g (95% CI)	p-value	
Full Fidelity Sample	High Stage 1	3,287 (272; 3,015)	0.37 (-6.89; 7.64)	+0.01 (-0.29; +0.32)	0.919	
	Stage 2	2,262 (243; 2,019)	-0.81 (-7.38; 5.77)	-0.03 (-0.31; +0.25)	0.810	
Complete Pairs Sample	Stage 1	584 (272; 312)	-3.77 (-8.94; 1.40)	-0.18 (-0.42; +0.07)	0.152	
	Stage 2	420 (243; 177)	-7.52 (-15.69; 0.65)	-0.35 (-0.73; +0.03)	0.071	

Medium/High-fidelity subsample (13 intervention schools)

Coefficient & Effect Size from stage 1 (KS1 maths) and stage 2 models (KS1 maths & CT test)						
Sample	Stage	n in model (intervention; control)	Coef (95% CI)	Hedges g (95% CI)	p-value	
Full Medium / High Fidelity Sample		3,665 (650; 3,015)	-3.14 (-7.84; 1.55)	-0.13 (-0.33; +0.07)	0.190	
		2,579 (560; 2,019)	-3.54 (-7.88; 0.81)	-0.15 (-0.33; +0.03)	0.111	
Complete Pairs Sample		1,349 (650; 699)	-1.10 (-5.42; 3.22)	-0.05 (-0.23; +0.14)	0.617	
		420 (560; 501)	-3.37 (-8.24; 1.50)	-0.14 (-0.35; +0.06)	0.175	

Appendix L: ScratchMaths team post PD evaluation

Post PD evaluation

The ScratchMaths team collected participant satisfaction data immediately after professional development events. The tables below present the summary findings of these across all seven hubs. The response rate was not provided.

Table 53: Y5 post PD feedback (percentage responses)

n=86	Excellent	Good	Satisfactory	Poor
How well did the PD course meet its aim of preparing you to teach the ScratchMaths Y5 curriculum?	67.7	19.4	4.3	0
How useful was the PD course in developing your classroom practice?	41.9	39.8	7.5	0
How confident are you that you will be able to teach the ScratchMaths content in your classroom?	30.1	52.7	9.7	0
How appropriate were the professional development materials?	65.6	23.7	2.2	0

Data source: Post-training satisfaction survey/ feedback forms'

Table 54: Y6 post PD feedback (percentage responses)

n=47	Excellent	Good	Satisfactory	Poor
How well did the PD course meet its aim of preparing you to teach the ScratchMaths Y5 curriculum?	54.4	22.8	1.8	1.8
How useful was the PD course in developing your Scratch programming skills?	57.9	19.3	1.8	1.8
How useful was the PD course in developing your classroom practice?	38.6	36.8	1.8	0

Data source: Post-training satisfaction survey/ feedback forms'

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.

OGL This information is licensed under the Open Government Licence v3.0. To view this licence, visit <http://www.nationalarchives.gov.uk/doc/open-government-licence/>

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 www.educationendowmentfoundation.org.uk



Education
Endowment
Foundation

The Education Endowment Foundation

9th Floor, Millbank Tower

21–24 Millbank

London

SW1P 4QP

www.educationendowmentfoundation.org.uk

Sheffield Hallam University

ScratchMaths: evaluation report and executive summary

BOYLAN, Mark <<http://orcid.org/0000-0002-8581-1886>>, DEMACK, Sean,
WOLSTENHOLME, Claire, REIDY, John and REANEY, Sarah

Available from the Sheffield Hallam University Research Archive (SHURA) at:

<http://shura.shu.ac.uk/23758/>

Copyright and re-use policy

Please visit <http://shura.shu.ac.uk/23758/> and
<http://shura.shu.ac.uk/information.html> for further details about copyright
and re-use permissions.