Education Endowment Foundation

INCREASING COMPETENCE AND CONFIDENCE IN ALGEBRA AND MULTIPLICATIVE STRUCTURES (ICCAMS)

Evaluation Report

December 2021

Maria Pampaka, Julian Williams, Jack Quinn, Diane Harris, David Swanson, Lawrence Wo, Abate Kenna, Graeme Hutcheson

MANCHESTER 1824
The University of Manchester

Education Endowment Foundation

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:

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

0207802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Department for Education | What Works Network | The Sutton Trust | Impetus Private Equity Foundation

Contents

About the evaluator

The project was independently evaluated by a team from the University of Manchester: Maria Pampaka, Julian Williams, Jack Quinn, Diane Harris, David Swanson, Lawrence Wo, Abate Kenna, and Graeme Hutcheson.

The lead evaluator was Maria Pampaka.

Contact details:

Dr Maria Pampaka

Ellen Wilkinson Building (B4.1)

Manchester Institute of Education

The University of Manchester

Manchester, M13 9PL

Tel: +44(0)1612757213

Email: maria.pampaka@manchester.ac.uk

Executive summary

The project

The Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS) intervention aimed to improve mathematics teaching and learning in Year 7 and Year 8 (pupils aged 11 to 13). Evidence from a nationwide survey of learners in 2008/2009 found that attainment in these areas has not improved since the 1970s, and for some groups of students it has fallen substantially (Hodgen et al., 2014). A team initially based at Nottingham University developed the intervention. A team from Durham University was responsible for the professional development of teachers throughout delivery of the intervention.

Teachers taught ICCAMS lessons during their regular timetable in the normal school day. The intervention consisted of a sequence of ten pairs of closely related lessons per year, making a total of 20 lessons per year. An associated mini assessment preceded each pair of lessons. The lessons were designed to help teachers use formative assessment in mathematics related to multiplicative reasoning and algebra. Activities are intended to be collaborative, set within contexts that are engaging for students, and use visual representations to help develop understanding.

Professional development took place at two levels. First, professional development leads (PD leads) trained two lead teachers per school at regional training days. After each training day, lead teachers were expected to teach the ICCAMS lessons and then conduct a one-hour cascade session for teachers in their schools. There were nine training days for lead teachers over the two years. The Durham University team worked with the PD leads during the pilot and supported them during the trial, as did the developer team. The National Centre for Excellence in Teaching Mathematics and Maths Hubs also supported the trial.

This was an effectiveness trial and involved 109 schools and 20,827 pupils. It had a two-arm cluster randomised design. Schools were randomly allocated to either use ICCAMS or to be part of the comparison group, which was expected to teach lessons as usual. This project started in August 2015 with a pilot phase and the intervention took place from September 2016 to July 2018. An implementation and process evaluation sought to assess fidelity and evaluate other aspects of the intervention using teacher and student surveys, observations of PD sessions, lesson observations, and interviews with stakeholders.

Table 1: Key conclusions

Key conclusions	
1.	Pupils in the ICCAMS schools made, on average, no additional progress in mathematics compared to pupils in the other schools. This result has a moderate to high security rating.
2.	Exploratory analysis suggests that there is no evidence that ICCAMS improved pupil progress in multiplicative reasoning or improved attitudes to mathematics compared to pupils in other schools but that pupils in schools that received ICCAMS did make the equivalent of one month's progress in algebra.
3.	Pupils eligible for free school meals in ICCAMS schools made the equivalent of one month's progress in mathematics and in the subscales of multiplication and algebra, on average, compared to equivalent pupils eligible for free school meals in the other schools. There was also some evidence of a more positive attitude to mathematics. These results may have lower security than the overall findings because of the smaller number of pupils.
4.	Teacher surveys found that 78% of lead teachers and 54% of cascade teachers said they were confident about ICCAMS teaching. Additionally, student and teacher surveys found some evidence that the intervention did change teachers' practice.
5.	One significant challenge was the cascade training. Only 55% of lead teachers reported managing all the expected cascade training sessions. In addition, although each cascade session was expected to be one hour, only 13% of teachers reported that the sessions were at least this length.

EEF security rating

These findings have a moderate to high security rating. This was an effectiveness trial, which tested whether the intervention worked at scale in a number of schools. This was a well-designed two-armed randomised controlled trial; 17.2% of the pupils who started the trial were not included the final analysis because their school did not provide test data. Implementation fidelity makes it harder to accurately estimate the size of the impact on the pupils in the trial.

Additional findings

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

The trial also compared pupil progress in multiplicative reasoning and in algebra. Pupils in schools receiving ICCAMS made no additional progress in multiplicative reasoning compared to pupils in control schools but did make one month of additional progress in algebra. The study also investigated any changes in students' attitudes towards maths and found no difference in attitudes between pupils in schools receiving ICCAMS compared to pupils in the control group.

The theory of the programme proposed four key steps: (1) lead teachers are effectively trained by PD leads, (2) lead teachers effectively train other teachers in their schools, (3) teachers then deliver ICCAMS lessons appropriately in schools, and (4) pupils receiving ICCAMS lessons achieve higher numeracy outcomes. The evaluation found evidence to support the first step but weaker or little evidence for the other steps.

Observations of the training judged that the training of PD leads was delivered well, with high levels of attendance and engagement. Lead teacher surveys found that over 80% of lead teachers felt that the PD was good or excellent, and 88% found the follow-up coaching visits from PD leads helpful. Regarding teaching, 78% of lead teachers felt confident teaching ICCAMS lessons and 85% said they would feel confident teaching ICCAMS again.

However, information from interviews and observations shows that the cascade sessions delivered by lead teachers to other teachers in their school were a significant challenge. Most lead teachers (55%) reported delivering fewer cascade sessions than expected. Those sessions were shorter than the programme expected. Only 13% of teachers in schools reported that cascade sessions were the expected length of at least an hour, with 76% of teachers reporting that the cascade sessions were 30 minutes or less. Factors that affected the amount of cascade time included teachers being busy with other responsibilities, the reluctance of some staff to engage with the approach, and staff turnover.

In turn, the quality and number of lessons taught varied in schools. The evaluation found that lead teachers were more engaged with the project principles and taught more lessons in the intended manner. Although this could have been due to training, lead teachers may also have been more experienced teachers who had more confidence in their capacity to adapt and manage change. According to teacher and student surveys, the intervention changed the classroom practice towards student collaboration and discussion. Changes in teaching practices were easier to observe in lead teachers. Other teachers changed their practice less and were less likely to be confident in delivering ICCAMS lessons. Other teachers also taught fewer ICCAMS lessons, with only 41% of surveyed teachers reporting that they taught all 20 lessons. As noted, the changes in teaching practice did not lead to consistent improvements to learner outcomes.

Cost

The average cost of intervention for one school was around £5,720, or £15 per pupil per year when averaged over three years. This estimate is based on the delivery of ICCAMS in Year 7 and Year 8 with, on average, 190 pupils receiving the intervention per school (as happened in this trial).

Impact

Table 2: Summary of impact on primary outcome(s)

Outcome/ Group	Effect size (95% confidence interval)	Estimated months' progress	EEF security rating	No. of pupils	P Value	EEF cost rating
Mathematics	0.04 (-0.07, 0.15)	0	3 padlocks	18,052	0.507	£ £ £ £ £
Mathematics, FSM eligible pupils	0.06 (-0.04, 0.16)	1	N/A	4,981	0.215	£ £ £ £ £

Introduction

Background

The ICCAMS mathematics intervention aims to raise engagement and attainment in mathematics at Year 7 and Year 8 by enabling teachers to implement formative assessment through a whole class teaching approach. The focus is on multiplicative reasoning and algebra, which cause particular problems for Key Stage 3 students. The programme, comprising 40 lesson plans, 20 associated mini-assessments, two paper tests, and 20 revisit activities, is intended to be implemented across a school, and is supported with a teacher professional development (PD) programme undertaken over nine days in two years. The lessons were designed to help teachers use formative assessment (and feedback) in mathematics, helping them to identify students' difficulties and misconceptions and how to address them. Activities were intended to be set in engaging contexts, collaborative when possible, and using visual representations to help develop understanding. This improved teaching was designed to result from the ICCAMS programme materials and teachers' PD and cascade PD in schools led by their lead teachers who were engaged in regional PD sessions.

The development of this intervention built on the findings of the initial ICCAMS funded by the Economic and Social Research Council (ESRC) (ICCAMS 1), which aimed to investigate ways of raising students' engagement and attainment by using formative assessment, which informs teaching and learning of mathematics in secondary school. The first phase of the study involved a longitudinal national survey of Year 9 students which used tests first developed in the 1970s under the framework of Concepts in Secondary Mathematics and Science (CSMS). This survey provided up-to-date empirical evidence on (1) the then current lower secondary students' understandings of, and difficulties with, algebra and multiplicative reasoning, (2) rates of progression across Key Stage 3 (KS3), and (3) differential performance across the cohort. In fact, the CSMS framework was influential in documenting the misconceptions of secondary students (Hart, 1981), which in turn motivated the developers of the initial intervention. This survey also enabled a comparison of students' understanding over the time since CSMS, which evidenced a decline in students' understanding in algebra and ratio since the 1970s (Hodgen, Coe, Brown and Kuchemann, 2014; Hodgen, Brown, Kutchemann and Coe, 2010).

In response to this, the 'ICCAMS 1 team' suggested the need for a more topic-focused formative assessment approach to the teaching of these areas (Hodgen et al., 2010). The ICCAMS approach was evaluated in the third phase of the original study through a matched controlled trial study with a group of 22 teachers and 600 Year 8 students. This intervention study showed that ICCAMS students made greater progress in attainment than a matched control group; this was considered a significant gain, the equivalent 'of about two years' normal progress in one year' (Hodgen et al., 2014, p.171). Table 3 overviews the key elements of the initial ICCAMS study compared to the current evaluation, which is detailed next.

Table 3: Comparison of key aspects of the previous evaluation of ICCAMS with the current evaluation

	Feature	Pilot to efficacy stage (from Hodgen et al., 2014)	Effectiveness stage (reported here)
Intervention	Intervention content	20 whole class assessment starters and 40 lessons (20 pairs).	Development of 4 new lesson pairs.
	Delivery model	Developer-led.	From developer-led to train-the-trainers cascade model of PD to all Year 7 and 8 teachers in school.
	Intervention duration	1 year (Year 8).	2 years (Year 7 and Year 8).
Evaluation	Eligibility criteria	Part of longitudinal ICCAMS study.	Mainstream English state secondary schools (or middle schools) with more than two class intakes for Year 7. Target: All Year 7 students.
	Level of randomisation	Not randomised: voluntary participation (22 Year 8 classes in 11 schools).	School (stratified in regions).
	Outcomes and baseline	Tests building on CSMS.	MALT Mathematics Test (and subscales)
	Control condition	Matched control (using propensity score matching from the broader longitudinal ICCAMS sample). Pre-post test.	Clustered randomised controlled trial (control/comparator: business as usual).

Intervention

During the first development/pilot year of the project, the ICCAMS maths intervention was adapted (1) for teaching over two years (rather than over one academic year as in the original project) and (2) to provide particular support for low attaining students and their classes. In addition, the University of Nottingham team developed material that explicitly describes the ICCAMS maths PD programme so that it can be delivered independently and with a degree of fidelity (including materials to train and support the Professional Development (PD) leads). The developers also extended the material from 20 lesson pairs to include further optional lesson plans for low and high attaining students.

In order to provide a comprehensive and transparent description of the ICCAMS, we utilise an adapted version of the Template for Intervention Description and Replication (TIDieR; Hoffmann et al., 2014), as per recommended reporting guidance (Humphrey, Lendrum, Ashworth, Frearson, Buck and Kerr, 2016), which occasionally will involve revisiting and overviewing what was extensively covered in the previous section. This description concludes with the logic model as agreed with the developers (and as presented in the protocol), which demonstrates the theorised processes by which the intervention inputs could lead to specified outcomes (Figure 4).

Brief name

Increasing Competence and Confidence in Algebra and Multiplicative Structures (ICCAMS)

Why (rationale/theory)

ICCAMS is the result of a research study (ICCAMS 1) which has shown promising results in raising students' mathematics learning, as noted earlier. The intervention aims to enable mathematics teachers to implement formative assessment, which was found in research to be an effective approach to increasing attainment and engagement (Black and Wiliam, 1998a, b; Black, Harrison, Hodgen, Marshall and Serret, 2011). Formative assessment, which refers to 'all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged' (Black and Wiliam, 1998, pp.7–8), was used in ICCAMS lessons to build on students' ideas, elicit and address misconceptions, encourage them to 'make sense' of mathematics, and to value their attempts to do so. ICCAMS was also developed to address critiques of formative assessment in regard to poor understanding of its implementation and that it needs to be integrated with subject-specific pedagogies (Bennett, 2011). ICCAMS was thus focused on algebra and multiplicative reasoning, which are topics known to create difficulties for this age group (for example, Hart, 1984; Hodgen et al., 2010) and the lessons are designed to expose dispositions through eliciting a range of student responses (for example, Ryan and Williams, 2007; Smith et al., 1994) so they can then be examined and developed through additional tasks and teacher-mediated discussion.

The approach places more emphasis on conceptual understanding and less emphasis on practice (Kilpatrick et al., 2001) and was intended as enrichment, rather than replacement, of a school's scheme of work for Year 7 and Year 8. In developing this mathematically-oriented approach to **formative assessment**, the developers of the ICCAMS approach were guided by the following key principles drawn from the research literature on mathematics teaching and learning:[1]

1. setting activities in **realistic contexts** (for example, Streefland, 1991); by 'realistic', the developers emphasised contexts that the students can imagine and engage with (so called 'experientially' real) rather than all activities being set in authentically real life contexts;
2. making **connections** between mathematical ideas (Askew, Brown, Rhodes, Johnson and Wiliam, 1997a);
3. encouraging **collaboration and talk** in mathematics teaching and learning (Slavin, Lake and Groff, 2009; Stein, Engle, Smith and Hughes, 2008); and
4. the use of **multiple representations,** such as the Cartesian graph or the double number line, to help students better communicate, understand, and connect mathematical ideas and to help teachers appreciate students' difficulties (for example, Gravemeijer, 1999).

---

[1] http://iccams-maths.org/our-approach/

The ICCAMS approach to PD is informed by the literature on teacher professional development (Adey, 2006; Adey, Hewitt, Hewitt and Landau, 2004; Cordingley, Bell, Evans and Firth, 2005) and by the same principles that guide the ICCAMS lessons. The design/structure of the PD involves teachers first engaging with each lesson and the associated mini-assessments followed by reflection on the lesson in the subsequent PD session after teaching the lesson in school.

Who (recipients)

All students in Year 7 of a school cohort, over a two-year period, up to the end of Year 8.

The mathematics teachers in these schools are also recipients of, or participants in, the PD (see also Implementers section).

What (materials)

Central to the ICCAMS mathematics intervention is a teacher handbook containing details of the intervention, lesson plans, the mini-assessments with brief commentaries, and guidance on adapting the intervention for low and high attaining students/classes. The handbook was provided to the teachers along with PowerPoint slides to use in their teaching and some GeoGebra files to be used in some lessons.

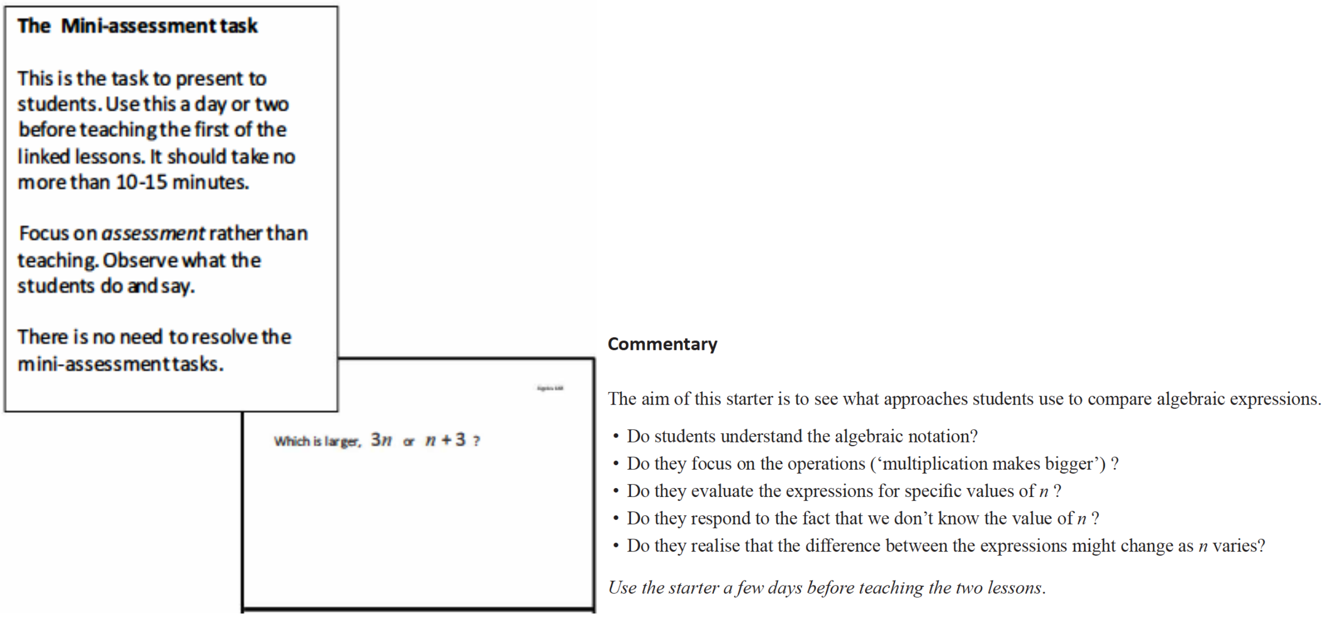
The intervention consists of 40 lessons organised in 20 pairs with associated (n = 20) mini-assessments and revisit tasks to be taught over the first two years of secondary school (Year 7 and Year 8). There are also four extra optional lessons provided to enable teachers to adapt the intervention for low and high attaining student/classes.

The two closely linked lessons that make up a pair are preceded by a related mini-assessment (10 to 15 minutes), as shown in Figure 1.

Figure 1: The intended sequence of ICCAMS lesson pairs and mini-assessments (adopted from ICCAMS protocol)

The mini-assessments (see an example in Figure 2) are intended to be used a day or two before teaching the first of the two linked pair of lessons and are designed to provide an opportunity for the teacher to observe students doing, and talking about, mathematics related to the lessons. During the mini-assessments, teachers are encouraged to gather information rather than teach or resolve the tasks.

Figure 2: Example picture from mini-assessment and revisit tasks (extracts from teacher handbook from slides presented to LD during PD sessions 1 and 2)



Each lesson plan consists of four pages (see Appendix 6 for an annotated description of lesson structure):

1. a lesson outline—providing a brief summary and intended as an aide-memoire whilst teaching;
2. a more detailed overview—giving the rationale for the lesson, background on the mathematical ideas addressed, and suggestions as to how the lesson could be adapted and followed up;
3. an extended annotated outline—intended to be read before teaching the lesson to help a teacher plan her/his interventions; and
4. further background—describing some of the key mathematical or pedagogical issues; this was also used in the PD.

An additional 61 potential revisit tasks were provided, each linked to a lesson. Revisit tasks are short (10 to 15 minutes) but designed to enable teachers to follow up and consolidate key ideas. Teachers were asked to choose at least ten revisit tasks to use with their class (five per year).

The order of lessons reflects the revised KS3 national curriculum and they are intended to be taught in the order presented in the handbook. Some optional lessons are provided that have been designed to enable teachers to adapt the intervention for low attaining students/classes and for high attaining students/classes (see example in Figure 6B in Appendix 6 for the flow of multiplication lessons).

Materials from the PD sessions in the form of PowerPoint files and handouts were distributed to teachers. Also provided were slides to help with the cascade training for other maths teachers in schools. Table 4 presents an example structure of a PD day with lead teachers, as adopted from the slides of one observed PD session.

Table 4: Outline of the first PD session

Time	Activity
09:45	Introductory activity
10:00	Reflection on lessons taught since last PD
11:30	*Break*
11:45	Reflection on mini-assessment
12:00	New lesson 1 outlined and discussed
13:00	*Lunch*
13:30	New lesson 2 outlined and discussed
14:30	Further planning, discussion of cascade training, adapting lessons for low attainers
15:15	Feedback, Q&A, evaluation
15:30	*Close*

Who (implementers)

The key people responsible for the implementation of ICCAMS are the PD leads who led the regional PD sessions, the teachers who attend these ('lead teachers'), and the other teachers in the school to whom ICCAMS should be cascaded (including all teachers of Year 7 and Year 8).

Lead teachers thus led cascade PD in their schools after attending the PD session led by PD leads and after teaching relevant ICCAMS lessons themselves. The other mathematics teachers (responsible for Years 7 and 8) then taught the lessons after their cascade PD sessions.

The professional development was provided for the lead teachers by five PD leads who were trained during the pilot/development stage of the project and were supported throughout by the ICCAMS developer and delivery teams.

The PD leads recruited to deliver the intervention have significant professional development experience in secondary mathematics education. The delivery team led a programme for the PD leads which involved a three-day session in May 2016, two days in September 2016, and ongoing single day conferences throughout the project. PD leads were also regularly networking with the developer team and each other through regular conference calls. PD leads were provided with presentation materials detailing tasks and discussion points, which they could adapt. Professionally produced videos were provided for discussion. Based on lesson extracts or clinical interviews, these videos are designed to exemplify students engaging with mathematics (including known misconceptions) and other aspects of the ICCAMS approach. PD leads visited each school in each year of the programme to observe teaching and provide feedback to the Lead Teachers.

The developer and delivery teams were working with the National Centre for Excellence in Teaching Mathematics (NCETM) in supporting the ICCAMS intervention and in particular with nine Maths Hubs across the five recruitment areas. In each area, the Maths Hub supported recruitment to the trial and continued promoting and supporting schools doing the intervention during the trial through the forming of an ICCAMS work group. The work group was led by an assistant PD lead who supported the PD lead during the project while developing the skills to lead the training after the trial. This assistant PD lead was nominated by the Maths Hub in each area, ideally located in a school not involved in the trial. They were expected to attend all the PD sessions for schools in their area, teach the ICCAMS lessons and run cascade sessions in their schools, and to support the PD lead in the PD sessions (becoming progressively more involved over the two years). There were also two days of support for the assistant PD lead, within the 'training the trainers' activities led by the delivery team.

What (procedures)

The PD model of this intervention involves two teachers from each school attending the regional PD sessions—teachers deemed suitably qualified by the school to lead and support the ICCAMS teaching. These teachers are then responsible for cascading ICCAMS to their colleagues who then are expected to use ICCAMs materials and approaches with all Year 7 pupils in the first year of the intervention and all Year 8 pupils in the second year. According to the information provided to schools before sign-up (see Appendix 3), 'At least one of the lead teachers should be senior in the maths department while the other can be any member of staff willing to attend and to disseminate the training back in school. Both teachers need to attend all nine PD sessions.' It was thus assumed these teachers would be experienced, amenable to the ideas of ICCAMS, capable of developing their own teaching, and capable of 'cascading' the training and supporting their colleagues.

Over the two years of the intervention there were nine whole-day PD sessions (six sessions in the first year, 2016/2017, and three in 2017/2018, see Figure 3) for lead teachers, which were organised and led by the PD lead in each of the five regional groups of around 20 teachers. During these sessions time was also allocated for discussion, planning, and reflection on the cascade PD. Schools (or mathematics departments/lead teachers) were responsible for identifying how to fit the ICCAMS lessons into their scheme of work.

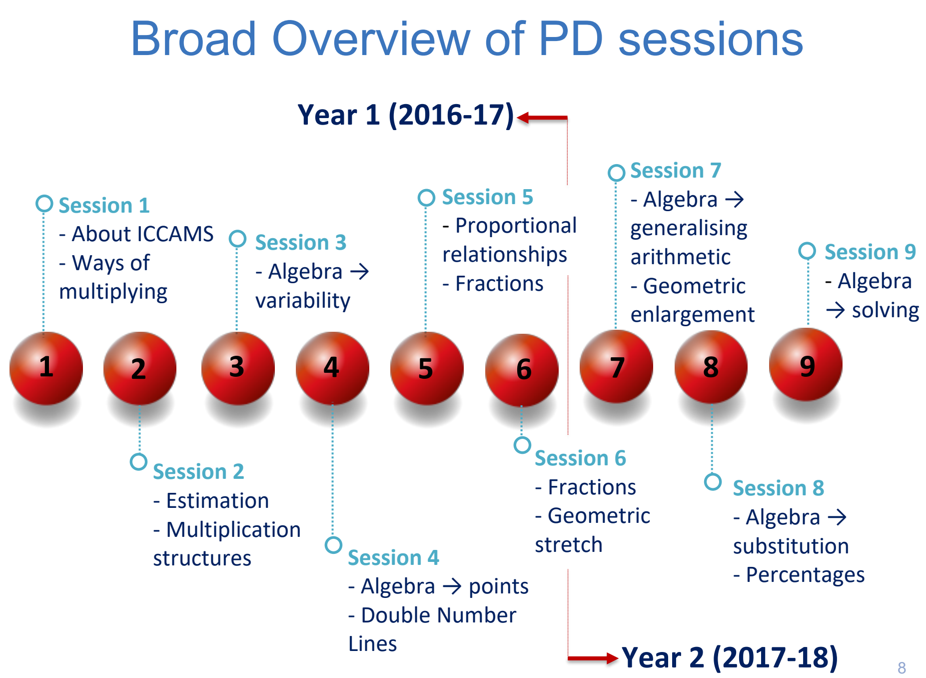


Figure 3: Slide from PD Session 1 showing the overview and timing of the PD sessions

Table 5 presents the main activities from the perspective of the lead teachers as specified by the developer and delivery team at the protocol (the full-time plan for the PD sessions is shown later in Table 6).

Table 5: Actions by lead teachers for planning and delivery of lessons during first year of intervention

Time frame	During PD sessions		At school
Year 1 Session one	• do (and extend) the mathematical tasks that the students will do • consider the possible difficulties students will have (including errors and misconceptions) • collaboratively plan how they will teach the lessons in their own classes	→ ↙	Teach the lesson [Mini-Assessment → Lesson 1 → Lesson 2]
Year 1 Session two to six	• reflect on the students' learning, the potential for its development in 'generalisation', the mathematics involved, the formative assessment strategies used, and how the lesson fits within the sequence of lessons • consider possible adaptations of the lesson • plan for how they will deliver the cascade training for those lessons they have already taught to the other teachers in the school	→	Deliver cascade training
Year 2 Sessions seven to nine	• as in Year 1 but combined • first session (PD7): planning (as Session one above) • second (PD8) and third (PD9): cover both elements of planning and reflecting	↔	Teach lessons Deliver cascade training

During the first year of the intervention, lead teachers teach the lessons and reflect on them at the next PD session before conducting cascade training and reflecting on the lessons with their department. However, during the second year, lead teachers conduct the lessons (see details about their sequence wth mini-assessments as well, in earlier section) then go on to teach cascade training immediately following this without the intermediary step of reflection at a PD session (but these lessons will then be reflected upon at the next PD session, as practised in year one). This was done so that in year one the PD trainers could model the sort of reflection needed in cascade training. Sets of cascade PD materials were provided to the lead teachers to use in their cascade PD for other teachers, including PowerPoint slides and video materials. Each cascade session which was also expected to cover issues addressed in the PD sessions, was designed to last around one hour, constituting a total of about eight hours over the two years of the programme.

How (implementation)

The ICCAMS lessons are taught by students' usual mathematics teachers during timetabled lessons in the normal school day. The agreement with schools (see also Appendix 3) asked those in the intervention group to adhere to the following guidelines:

- teach all lessons to all classes;
- conduct mini-assessments;
- two lead teachers per school to attend external PD regionally (for nine full days across the two years); and
- lead teachers to organise nine, hour-long cascade training sessions (one after each of the nine PD training sessions) to teachers in their department that did not attend the external PD sessions.

Where (setting)

The ICCAMS is implemented on-site in participating schools by teachers. ICCAMS lessons are taught in students' normal mathematics classroom (equipped with a data projector).

PD sessions, during the project, took place in five different locations, given the five regions participating. Venues were school sites or community spaces, which were booked by those running the trial. To find venues, mid-points between participating schools were found, accounting for the often large distances between schools in the same region.

When and how much (dosage)

The 'expected' ICCAMS offer for students consists of 40 lessons (50 to 60 minutes each), 20 associated mini-assessments (15 minutes each) and at least ten 'revisit' tasks (15 minutes each) to be taught over the two-year period as shown in Table 6. For students this implies an engagement of about 40 to 50 hours with ICCAMS over the two years.

Additionally, a short formative test is administered at the beginning of both Year 7 and Year 8.

Table 6: Overview of the intervention PD training and lessons to be taught

Timeframe	Professional development	Cascade sessions	Lesson IDs by lead teachers	Lesson IDs by cascade teachers
Year 1, 2016/2017—Year 7				
October	PD1	Cascade 1	1	
November	PD2	Cascade 2	2 and 3	1
December	PD3	Cascade 3	4	2
January	PD4	Cascade 4	5	3
February				4
March	PD5	Cascade 5	6 and 7	5
April	PD6			
May to July		Cascade 6	8 to 11	6 to 11
Year 2, 2017/2018—Year 8				
September	PD7			
October		Cascade 7	12 to 14	
November	PD8			12 to 14
December		Cascade 8	15 to 17	
January				15
February				16 and 17
March	PD9			
April to June		Cascade 9	17 to end	17 to 20 end

Tailoring

Teachers were encouraged to adapt ICCAMS in several ways (aligned with formative assessment):

- The mini-assessments were designed to enable teachers to predict what difficulties different students would encounter and how students would engage with the lesson tasks in order to help teacher better respond to students (and mediate class discussions).
- Explicit guidance was provided in the lesson materials on ways of adapting or following up on the lesson. The guidance emphasised that it is not essential to complete all the tasks in each lesson, stressing that it is more important for students to engage with the relevant mathematics than they complete all the tasks.
- Additional lesson options were provided for tailoring the programme towards 'low' or 'high' attaining classes.
- Teachers were asked to choose at least five revisit tasks to use each year on the basis of their assessment of students.

Beyond these options given to teachers for variation and adapting the lessons, a consistent point was that the developers did not want schools to tailor the intervention (but instead advised teaching the lessons as they are).

Modifications

No modifications were reported during the trial. We leave a discussion of that for the results section.

How well (planned)

Effective implementation requires all mathematics teachers to receive cascade PD before they teach relevant lessons (delivered by lead teachers after they themselves have received ICCAMS PD and taught the relevant lessons). Hence, schools (or their mathematics departments) need to ensure that time is allocated for the cascade PD. In addition, at least one of the two lead teachers should be a 'senior' teacher in their school mathematics department and, thus, sufficiently experienced to lead cascade PD and with sufficient authority to advocate for the allocation of time for the cascade PD within school.

Figure 4 presents the logic model agreed with the developer and delivery teams during the protocol stage.

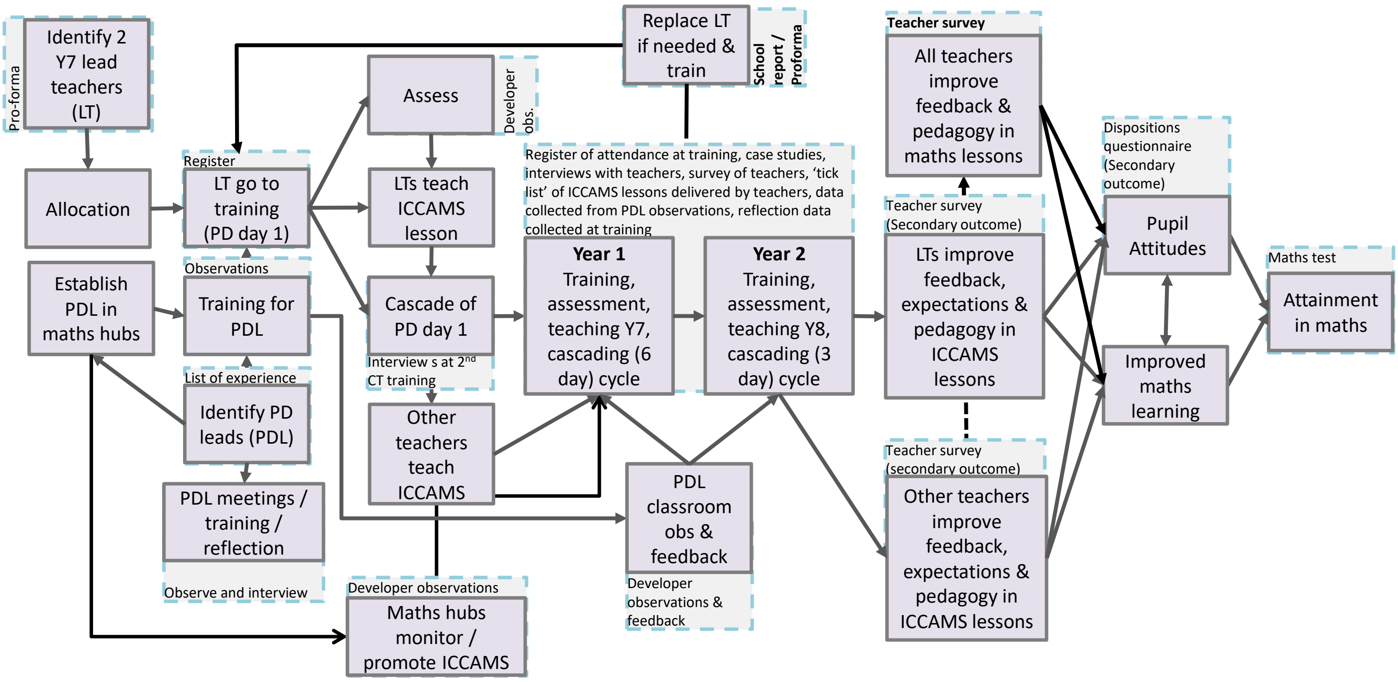


Figure 4: Logic model for the ICCAMS—as presented in the evaluation protocol

If all were to go according to plan, the logic of the intervention would include the following main steps, each of which might be considered to engage teachers and learners in a dialogue.

(i) The design of the lessons (which has been tested in previous research and piloting) matches the curriculum requirements and so the testing of the learning outcomes for the national curriculum for mathematics (including a strong multiplicative reasoning and algebra component) and the lead professional developers (lead PDs) have worked on the design to prepare their programme of support for teachers appropriately in light of the above.

(ii) Then the selected lead teachers from each school work together in groups with their PD leads to ready these lead teachers to teach the ICCAMS lessons to their own classes and to lead their colleagues in their departments to follow them similarly.

(iii) The cascade teachers are led in the school professional development programme by their own school's lead teachers to understand the lessons and teach their classes appropriately.

(iv) The result is that ICCAMS lessons are taught appropriately in the classrooms offering these children a series of ICCAMS tasks and 'learning experiences' including formative assessment (FA), task engagement, dialogue, and reflection and metacognition.

(v) Teachers' teaching practice is expected to improve, at least in the way they use FA and engage with classroom discussion.

(vi) This causes better learning, more engagement, and higher achievement or attainment than would otherwise have occurred, which is made visible in tests of mathematics and attitudes.

Before introducing the methods of this evaluation, it is important to understand the background research that informed ICCAMS and its various elements and provide an overview of recent empirical evidence that emerged after the conception and development of the programme (that is, including the last ten years). This is necessary in order to contextualise the findings of this evaluation and provide a theoretical grounding for the emerging evidence base. This, therefore, involves a brief excursion into the broader literature that led to the ICCAMS intervention design as well as similar studies or interventions that may have run in control schools in parallel with the intervention.

The question at issue is this: how are interventions like ICCAMS considered to 'work' (or not) in terms of improving teaching and—especially—learning opportunities and learning outcomes? The reasons for the outcomes of this particular intervention can only be understood within the context of the rationale for this intervention—why, how and when such interventions might work—and so we begin with what motivated this intervention design in the first place.

Therefore, in the next sections we turn to the policy context, then to the research literature.

Situating ICCAMS within the current wider policy and practice landscape

ICCAMS directly addresses concerns about the development of mathematical performance of students in lower secondary school mathematics, which is of concern to many stakeholders, but it does so in light of the understandings in the educational field of the conceptual developments required for learners to progress beyond immediate performance to future development as mathematicians. Successful performance with multiplicative structures and algebra provide a student with the majority of what is required on test scores at GCSE, but also understanding of the same is thought to provide what is needed for progression in mathematics thereafter. As such, this topic signals a nexus between policy imperatives and educational praxis—perhaps a fortuitous connection.

Starting from formative assessment, the most recent report in 2015, the 'Final Report of the Commission on Assessment without Levels',[2] stressed the importance of a much greater emphasis on FA as an integral part of teaching and learning that would have multiple benefits such as improving teaching, raising standards, and reinforcing schools' independence to adapt teaching to fit not only the needs of the pupils but also the strengths of the staff. However, the interest and efforts on this front are not new.

Assessment for Learning (AfL) is synonymous with formative assessment (Wiliam, 2011; Wiliam and Thompson, 2007), which was highly significant in the antecedents and formulation of ICCAMS. AfL is assessment that is designed to inform teaching and learning, takes place continuously during classroom teaching, and can be used to find out where the students are and revise the next steps in their learning hence informing the planning of teaching. As such, AfL is one of the most significant reasons for assessment in the classroom (Assessment Reform Group, 2002). In practice, the ICCAMS project was concerned with implementing the FA/AfL approach in a particular context, that is, the mathematics classroom for Years 7 and 8 and, in particular, for the topic areas of multiplicative and algebraic structures.

In order to support teachers with assessment, Assessing Pupils Progress (APP), produced by the Qualifications and Curriculum Authority (QCA, 2010), provided a structured approach to periodic assessment with the intention of enabling teachers to use diagnostic information about students' strengths and weaknesses and then to track students' progress as they progressed through the national curriculum. The scheme included guidance for formative assessment throughout but also summative assessment at the end of each key stage. The Department for Children, Schools and Families (DCSF, 2008)[3] also offered guidance[4] in the form of a handbook for Key Stage 3 teachers, which sets out key features of the three 'assessment viewpoints' (p.3), that is, (1) day-to-day (for example, explicit learning objectives shared with pupils and immediate feedback), (2) periodic (for example, a broader view of progress across subjects, the use of national standards in the classroom, and improvements to medium-term curriculum planning), and (3) transitional (for example, formal recognition of achievement, reports to parents/carers and next teacher(s), and the optional use of external tests).

A pilot project to assess this initiative ('Making Good Progress', DCSF, 2011)[3] found that although the initiative increased teachers' accountability and the accuracy and consistency of assessment practice, the materials did not necessarily improve teachers' evaluation of pupils' progress in lessons. It was believed that this was because other key teaching skills were needed, for example, those involving questioning and listening. Drawing on such results, in this evaluation we attempt to also explore the mediating effect of teaching practices.

---

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/483058/Commission_on_Assessment_Without_Levels_-_report.pdf

[3] The impact of the 'Assessing pupils' progress' initiative April 2011, No 100226 https://www.gov.uk/government/publications/the-impact-of-the-assessing-pupils-progress-initiative

[4] Assessing pupils' progress in mathematics at Key Stage 3: Teachers' handbook. Nottingham: DCSF Publications https://dera.ioe.ac.uk/14817/7/The%20new%20APP%20handook_Redacted.pdf

This brings us to the topic of professional development more generally and the expectations for and of teachers in this regard. The expectations of the Department for Education (DfE) for professional development (PD) are set out in 'The Teachers' Standards' (Department for Education, 2016). This publication—which draws heavily on the findings of the most recent (at the time) review of evidence on effective teacher PD: Cordingley, Higgins, Greany, Buckler, Coles-Jordan, Crisp, Saunders and Coe, (2015)—reports that PD aiming to modify teachers' practice was most effective when it included collaborative activities that support the intended learning outcomes for the students. The document also distinguished between PD programmes and short activities (for example, one-day, stand-alone activities) with the latter not likely to have sustained impact on students' learning outcomes. They argue that effective PD must be prioritised by school leadership and when effective it 'builds-in peer support for problem-solving; includes focused discussion about practice and supporting groups of pupils with similar needs; challenges existing practice, by raising expectations and bringing in new perspectives; and includes support from someone in a coaching and/or mentoring role to provide [teachers with] modelling and challenge' (Department for Education, 2016, p.9).

Official reports on teachers' surveys also provide some evidence of how teachers perceive PD opportunities in practice. For instance, in TALIS-2018[5] it was reported that overall scores for collaboration are higher on average among primary teachers than among lower secondary teachers (with effect size = 0.34), which reflects a greater frequency of collaborative activities, including professional development in primary schools.

In regard to mathematics, in 2006, the DfE funded the National Centre for Excellence in the Teaching of Mathematics (NCETM), which aims to raise levels of achievement in mathematics and to increase students' appreciation of mathematics. To achieve this, the NCETM aimed to provide easier access to high quality, evidence-based, mathematics-specific continuing professional development (CPD) for teachers at every point of their careers. Their 'Maths Hubs' programmes have been the key means of support for teachers covering all school and college phases, from early years to post-16. In Key Stages 1, 2, and 3, the majority of their support was via their Teaching for Mastery programmes. The mathematics Teaching for Mastery, according to the official government website,[6] 'aims to empower and equip teachers to improve pupils' understanding and attainment in maths'. Like ICCAMS, the Mastery Programme included professional development sessions in its approach to teaching mathematics, support with subject knowledge, and provided classroom resources. At KS3, the NCETM provide Mastery resources[7] for key themes in the mathematics curriculum. Of particular relevance to ICCAMS are:

- multiplicative structures—connecting numbers by repeated addition, array models, and various other models of multiplication, which, at KS3, relates to ratio, proportion, rates of change, percentage increase or decrease, enlargements, and other geometrical relations and transformations and so similarity and trigonometric ratios; and
- algebraic structures—generalisation of number patterns and relations, simplifying and manipulating expressions, and equations and formulae.

However, there are significant differences between ICCAMS and Mastery in the scale of the two projects that are relevant to contextualising the current evaluation within the broader policy and practice context: Mastery was a national programme that had a wide remit with some internal inconsistencies which the NCETM hubs had to work out in their own ways, while the ICCAMS was relatively focused and centrally driven based on a strong perspective in theory, and implemented more centrally in practice.

Situating ICCAMS evaluation within the current research landscape

The ICCAMS builds on a base of research about formative assessment and PD that is widely published internationally and also well known and used in many professional contexts, though there is a diversity of practices that claim to be derived from the findings of these researches. The evidence for the ICCAMS development and its principles (as noted

[5] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/873922/Teaching_and_Learning_International_Survey_2018_March_2020.pdf
[6] https://www.gov.uk/guidance/join-the-maths-teaching-for-mastery-programme
[7] https://nmmathshub.co.uk/NCETM-KS3-Mastery-Resources/

earlier) is also well established and covered extensively by the developers in their previous work and their own justification for this intervention (for example, Hodgen et al., 2014). Our aim here is not to replicate such a review, which precedes the design of such an intervention, but rather to situate the evaluation design, including our approach to the 'process' aspects of the evaluation, and the results of the evaluation within the broader context of research in mathematics education and PD. The intention is to understand in what ways and in what conditions the development of a FA PD programme might be effective in developing pedagogy and in improving learning in the context of this evaluation and the emerging results.

We present here an overview of this literature, which is based on a broader review of the literature about FA and PD in general (see methodological details of our approach to this review in Appendix 1),[8] while focusing on the aspects pertinent to ICCAMS. We start from what works for efficient PD in mathematics education, followed by an overview of the effectiveness of FA, and integrated FA with PD initiatives. We then focus on the role of teachers on such attempts along with teacher-related impacts before we move to the student related outcomes, including attainment and emotions, and with a lens on multiplicative reasoning and algebra.

What is found to work for 'efficient' professional development in secondary mathematics?

There is widespread acknowledgement of the teachers' crucial role and also agreement that high quality[9] teachers are essential for the successful teaching and learning of mathematics (Chval, Abell, Pareja, Musikul and Ritzka, 2008; Lomibao, 2016) with professional development as a crucial enabler of such desirable improvements (Carlson, 2002; Mokhele and Jita, 2012). Teachers, thus, can benefit from PD by keeping up-to-date with new developments in teaching and learning to ensure their practice is continually developed and improved, which effectively will benefit student learning (Wiliam, 2016). Understandably, PD programmes need to be well-designed and well-implemented to be effective (Darling-Hammond, Hyler and Gardner, 2017).

With varying degrees of success, PD is undertaken worldwide with commonalities between the various initiatives reported, for example, conditions such as the location of the schools (rural, urban) and the socioeconomic factors at play within the schools themselves and their communities (Pournara, Hodgen, Adler and Pillay, 2015; Zakaria and Daud, 2009; Wong, 2010; Watson and Beswick, 2011; Tytler, Symington, Darby, Malcolm and Kirkwood, 2011; Goos, Dole and Geiger, 2011). These studies cover the need for PD to deal with low achievement groups, assessment including diagnostic tasks, and the need for collaborative interactions with colleagues. This evidence from the international literature should be viewed as an illustration of the breadth of foci in studies involving PD, and indicating some of the variation in approaches as also shown in the review by Holmqvist (2017) of the most frequently cited (according to English Google Scholar) articles on collaborative PD for mathematics teachers. The review showed that PD can inform practice-based research and contribute to the development of classroom-based teaching and learning.

In what follows we focus on issues of direct concern to this intervention's evaluation, starting from the importance of peer support networks, the role of teacher reflection, and the optimal duration of effective PDs. Later sections will deal with possible implications for developing teachers and teaching, and thence to the consequent effects on learners.

The importance of peer support networks and reflective practice for professional development

Peer-networking can be realised through various models as illustrated in Holmqvist's (2017) review, which identified five PD models: lesson study (Takahashi and McDougal, 2016; Takahashi, 2011; Cajkler, Wood, Norton and Pedder, 2014; Warwick, Vrikki, Færøyvik Karlsen, Dudley and Vermunt, 2019; Warwick, Vrikki, Vermunt, Mercer and van Halem, 2016; Archer, Morgan and Swanson, 2021), educational action research, teaching research groups, educational design research, and learning study. Whether mentoring or even coaching, the literature puts considerable weight on teachers working together to include shared experiences or observations of actual lessons where children's learning is in focus, and evidence from the classroom becomes germane for teachers' discussions. It is thus important to consider evidence on how teacher reflection on such experiences can affect the impact of PD.

---

[8] See www.teleprism.com/iccams-evaluation/LR.pdf
[9] It should be noted however, that there is less agreement as to what 'high quality teaching' might involve

For example, the secondary school teachers in the small scale study of Cajkler et al. (2014) reported that lesson study helped them to understand their students, to develop less teacher-centred approaches, and to have a stronger sense of teacher community. However, the research also found that there were substantive organisational challenges if the use of lesson study in the school were to expand (for example, lesson study can be resource intensive). An adaptation of lesson study, the learning study of Lau and Yuen (2013), draws on the 'theory of variation' to explain how a learner might come to understand or experience a given concept in a particular way (see also Marton and Säljö, 1976). This strategy enables teachers to reflect proactively on their teaching, to organise their lessons, and to manage student differences in learning mathematics. By identifying students' misconceptions, teachers are better placed to understand teaching from their students' perspective, which in turn helps teachers to develop pedagogical content knowledge in mathematics. Indeed, the variations in children's responses to diagnostic assessment tasks is a common focus for lesson study around the world and would fit with the intentions of ICCAMS mini-assessments very neatly if PD were developed this way in the cascading process.

Studies also stressed the need for strong leadership and support from senior managers in addition to time and resources (McNeill, Butt and Armstrong, 2016) for sustainable engagement with particular programmes (Kale and Selmer, 2014; Baker, Gersten, Dimino and Griffiths, 2004). The literature also suggests that a collaborative approach for networking and sharing experience amongst peers is preferred to the top-down model. For example, the collaborative approach of McNeill et al. (2016) was described as being more successful and more easily accepted than the 'delivery' training approaches or top-down cascade models and there were early indications of improvements in student performance at the participants' schools and colleges adopting 'collaboration'. The top-down model has also been criticised because it suffers information dilution the further it is from the original source (Lomibao, 2016) and because the model generally reflects a one-time and one-size-fits-all approach (Kale and Selmer, 2014).

However, peer networking is not only dependent on teachers' willingness to spread their knowledge but also on the willingness of other teachers to learn from their peers (Hatch and Lee, 2010; Tirosh, Tsamir and Levenson, 2015). A survey of 9,000 U.S. teachers over two years showed that teachers improve faster when they are in schools with better quality collaboration, where better achievement gains for students are also observed (Ronfeldt, Farmer, McQueen and Grissom, 2015).

Reflection is widely presented as the prime enabler of PD: teachers thereby self-evaluate their professional practice and gain a greater understanding of the broader contexts of teaching and learning. This aspect is so significant that PD projects focus on teacher reflection by providing teachers with opportunities to focus their attention on the mathematical activity and learning outcomes and consequently analyse and improve specific aspects of their teaching (Parada and Pluvinage, 2014).

The literature often suggests that focused discussion enables teachers to reflect on their own practice and develop a deeper understanding of teaching and learning, and that this is an essential step to improving practice. Reflection can enable teachers at all stages of the profession to examine their successes, struggles, and failures and consider options for change that will impact student learning. Internet technologies have also been used to facilitate such discussions.

How long should professional development last to be effective?

In the mathematics education literature, a number of traditional narrative literature reviews suggest that the duration and intensity of PD are crucial: in particular, an extended programme of two years is necessary for significant and sustained professional change to take place (Adey et al., 2004). Whilst this finding aligns with the views of experts in the field (for example, NCETM, 2009), the hard 'causal' evidence base is extremely limited. Time is indeed an issue for such programmes even though there is less agreement on the optimum duration: more time enables regular opportunities for reflecting on teaching practice to cultivate the development of teacher knowledge (Kale and Selmer, 2014) and can result in more impact (Fletcher-Wood and Zuccollo, 2020). Related to that, there is evidence that changes in teachers' views about lesson outcomes, teachers' efficacy in implementing new ideas, and willingness to give up their current teaching practices all develop at different rates (Witterholt et al., 2016). Other studies suggested that it may be the repetition of specific practice skills in PD that is important for effects on student outcomes rather than the length of the programme (Basma and Savage, 2018; Kraft, Blazar and Hogan, 2018).

A systematic review of the effects of PD on mathematics teachers (Gersten, Taylor, Keys, Rolfhus and Newman-Gonchar, 2014) identified 643 studies of PD relating to school mathematics, although only five met the What Works

Clearinghouse evidence standards.[10] All these five studies involved significant PD contact time yet only two reported positive effects on learners' attainment. Therefore, whilst extended contact time may be necessary, it is not necessarily sufficient. Ruthven et al. (2017) also proposed that the scale and quality of PD provided to teachers (in the U.K.) ought to be more explicitly identified as a potentially crucial variable in meta-analyses of research in this area. Their project suggested that limited commitment to PD at both system and school levels may be a major factor inhibiting the successful implementation of potentially more effective teaching practices that call for substantial professional learning.

In sum, we conclude that the duration required for PD to achieve outcomes is not known but that professionals judge that any significant change does take place generally on long timescales: to this should be added that it is also a matter of the quality of collaboration and reflection that is afforded within this time that might be considered important.

What these studies add up to is the notion that PD must not only be aligned with professionals' perceptions of their needs and their previous experience but informed by expertise and research, and that significant development must be social, based on work with peers over substantial periods of time: significant change is slow and expensive (but then, it can be argued, ineffective PD is more expensive). Next we review FA initiatives within PD and their effectiveness.

Effectiveness of formative assessment initiatives within professional development

Formative assessment is a flexible and informal way of assessing students' progress and their understanding of a particular subject matter. It can help teachers to recognise where their students are struggling so they can address problems immediately and, as such, is highly important in all teaching and learning situations. Although FA is not currently explicitly included in the EEF Toolkit, three key elements of FA are amongst the approaches with the highest impact: feedback, peer tutoring, and metacognition/self-regulation. However, the EEF Toolkit[11] does report that 'educational (rather than psychological or theoretical) studies tend to identify positive benefits where the aim of feedback is to improve learning outcomes in reading or mathematics or in recall of information' (EEF Toolkit, p.16). In this section we overview first the pre-history related to the ICCAMS FA approach and then recent broader evidence of similar PD/FA initiatives.

Various initiatives and projects on high quality instruction built on the earlier programme of Cognitively Guided Instruction (CGI) PD, which, according to Carpenter, Fennema, Franke, Levi and Empson (1999) was based on an integrated research programme with the following foci:

> *'(a) the development of students' mathematical thinking; (b) instruction that influences that development; (c) teachers knowledge and beliefs that influence their instructional practices; and (d) the way that teachers' knowledge, beliefs, and practices are influenced by their understanding of students' mathematical thinking' (p.4).*

The teachers in the CGI study changed for many reasons but the above study reported two as critical:

> *'(a) The teachers learned the specific research-based model that formed the basis of the teacher development program, and (b) the teachers used that model in the classroom. The research-based model served as a catalyst between teachers' intuitive knowledge and principled knowledge of their own students' thinking, which the teachers developed as they taught' (p.431).*

CGI was also used in the context of other studies, such as one (Franke, Webb, Chan, Ing, Freund and Battey, 2009) which found that although teachers freely ask their students questions, they find it much more difficult to know how to follow up on their students' ideas, that is, teachers struggle to elicit further information to help them understand their students' initial explanations. This is why the organisation of plans around quite specific assessment items with known variations in students' responses, such as those in ICCAMS, might be so important to teaching and learning.

Another project of note is the Mathematics Teacher Development Project (of Simon et al., 2000) which in turn aligns with the Realistic Mathematics Education of Gravemejer, Streefland and Freudenthal (for example, Gravemeijer, McClain and Stephan, 1999; Streefland, 1991). This is also aligned with Visnovska and Cobb's (2015) study in the U.S.

[10] https://ies.ed.gov/ncee/wwc/ It should be noted that the WWC criteria sometimes do seem overly strict: see also https://www.alliance4usefulevidence.org/assets/Alliance-FUE-reviews-booklet-3.pdf
[11] https://educationendowmentfoundation.org.uk/public/files/Toolkit/complete/EEF-Teaching-Learning-Toolkit-October-2018.pdf

(a five-year PD design experiment) where they found that 'leveraging teachers' existing practices and concerns was important in supporting them to focus on students' mathematical reasoning and develop adaptive practices' (p.1233).

The Cognitive Acceleration through Mathematics Education (CAME) project, another initiative influenced by the CSMS framework as is ICCAMS, followed the success of Cognitive Acceleration through Science Education (CASE) and aimed to improve children's thinking and achievement in mathematics (Goulding, 2002). A remarkable feature of the CASE evaluation (Adey and Shayer, 1994) was the finding that (1) it impacted significantly on GCSE exam results several years later and (2) the impact on GCSE grades applied beyond science, right across the curriculum. The CAME programme, developed at King's College and later implemented in a number of places round the world, used the same principles as the science project and also followed a similar format, that is, one lesson per two weeks for two years in Years 7 and 8 (12- to 14-year-olds) in secondary school. Cognitive acceleration depends on three core principles: (1) 'cognitive conflict' (from Piaget), that is, the necessity for students to develop by accommodating existing schemas (thus actively constructing knowledge) rather than merely assimilating or 'taking in' information passively (for example, from transmissionist, teacher-centred delivery), (b) 'social mediation' (from Vygotsky)—encouraging dialogue and discussion—and (c) metacognition, that is, the opportunity for students to reflect on their own learning in the cognitive acceleration lessons, which helps to consolidate what has been learned.

There were favourable post-test and long term national examination effects in mathematics with CAME, however as a methodology it is 'unfamiliar to many, believed by others to be permanently discredited, and is forgotten by some' (Shayer and Adhami, 2007, p.1). Shayer and Adhami (2007) used this approach with 2,500 11- to 13-year-olds which produced a large (0.8 SD) long term effect on the achievement of students by the time they reached 16. More recently, Finau, Treagust, Won and Chandrasegaran (2018) introduced their own cognitive acceleration programme in mathematics with Year 8 students (in an experimental design) which provided evidence suggesting that learning mathematics under the CAME programme could have a positive effect on levels of students' self-regulation, motivation, and mathematics achievement.

This is a persuasive pre-history to the ICCAMS approach and one that is based in the wider literature on Formative Assessment. An overview of key findings from the literature reviewed concerning what is known about the effectiveness of various approaches to FA and how this depends on the development of appropriate pedagogical implementations of FA through various PD strategies is presented below (see also Table 1.A in Appendix 1):

- FA can promote learning but the range of effect sizes varies between what Wiliam (2007b) has called 'weak' (ES = 0.16) and 'strong' (ES > 0.5) forms of metacognitive classroom implementation of FA (see the EEF Toolkit).
- Weaker forms of FA practice involve little more than generic routines (Marshall and Drummond, 2006).
- Stronger forms of FA involve feedback directly relevant to the task and engaging the learner in metacognitive reflection and identifying new work to address the next steps in learning.
- This makes significant demands on the teacher's understanding of the mathematics and especially potential learner trajectories through particular knowledge domains (for example, knowledge of difficulties and misconceptions and how to use these to effect conceptual change) (Smith and Gorard, 2005; Watson, 2006; Bennett, 2011; Shulman, 1986, 1987).
- The use of grades and scores in feedback (such as in APP) (Ardron and Monahan, 2010; Slade, 2009) tends to decrease or cancel the positive effects of FA as it can draw attention away from the learners' focus on their conceptual understanding and what they need to learn next (Hume and Coll, 2009; Li, Klahr and Siler, 2006).
- Peer formative assessment feedback and discussion can be as effective as learner-teacher dialogues (Ding and Harskamp, 2011; Davis, Kumtepe and Aydeniz, 2007).
- CPD programmes focusing on FA had mixed results on teaching with evidence that the most effective involved small teacher groups engaging in action and reflection or inquiry/practitioner-research, were extensive over time, and made use of well researched and designed materials introduced by experts (Wylie and Lyon, 2015; Yin, Olson, Olson, Solvin and Brandon, 2015; Kramarski, 2009; Kramarski and Revach, 2009).
- The evaluation of a recent EEF-funded relevant intervention (Embedding Formative Assessment)[12] for Key Stage 4 showed that students in schools that followed this whole-school approach made about two months' additional progress in their Attainment 8 GCSE scores; no evidence was found, however, on improved maths GCSE (or English) attainment specifically.

[12] https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/embedding-formative-assessment/

- FA is considered to be an important part of ‘connectionist teaching’ and is antithetical to ‘transmissionist’ teaching or ‘delivery’ teaching of procedures to prepare for test performance (Askew et al., 1997a; Askew, Brown, Rhodes, Wiliam and Johnson, 1997b; Swan, 2006; Williams, Black, Davis, Hernandez-Martinez, Hutcheson, Nicholson, Pampaka and Wake, 2008).

What is needed, therefore, is mathematical and task-situated formative assessment practices: this is what ICCAMS aimed to provide. Let us not forget the metacognitive, however, which involves discussion around the focal mathematical ideas, but, crucially, also opportunities for learners to reflect on their progress and further learning needs (in this respect we include teachers as learners in PD also). We turn to this next.

Teacher-related factors and impacts within professional development and formative assessment initiatives

There is agreement on the influential role of teachers on students’ learning (Lomibao, 2016) and the need for high quality teachers for the successful teaching and improved learning of mathematics (Chval et al., 2008; Foster, Toma and Troske, 2013) but there is less agreement as to what ‘high quality teaching’ involves.

In fact many studies, especially in the field of educational effectiveness research, attributed a lot of the variance of students outcomes to the class level and in particular to the individual **teacher quality,** and to practices in their classes (Muijs and Reynolds, 2011). The effects of teacher characteristics, such as level of education and teaching experience, upon student outcomes, has also been the focus of many studies (see review of Burroughs, Gardner, Lee, Guo, Touitou, Jansen and Schmidt, 2019). In mathematics, Rockoff (2004) found a positive relationship between teacher experience and student achievement but that this plateaued after only two years of teaching. Studies have also reported the relationship between teachers’ characteristics and their participation in PD programmes (for example, Akiba, 2012) including lesson study (Lomibao, 2016). Another (RCT) study in Italy (Argentin, Pennisi, Vidoni, Abbiati and Caputo, 2014), even though it did not find evidence of PD improving mathematics achievement, detected some heterogeneity in relation to the age of teachers (that is, the treatment appeared effective for middle-aged teachers). In a recent analysis examining the consistency of teacher characteristics and their relationship with student outcomes in the Trends in international Mathematics and Science Study (TIMSS) of five countries (including England), little consistency was found except only regarding instructional quality (Blömeke and Olsen, 2019).

Muijs, Kyriakides, Van der Werf, Creemers, Timperley and Earl (2014) warn about the lack of consideration of PD in effectiveness research and highlight the need for such studies not only to identify effective practice, but also to change practice. From the perspective of PD evaluation studies there is also some recent focus on the influential role of **teacher identity** for the success of PD (Skott, 2019; Gresalfi and Cobb, 2011) but also, and crucially, on how they can re-adjust (or ‘re-author’) these identities to change their practice after participating in PD (Darragh and Radovic, 2019).

The literature on PD and FA is actually overwhelmed by evidence on how teachers impact the success or not of various interventions, but also on the effects of such initiatives on teachers’ attitudes, beliefs, and also practices (for more detailed descriptions see www.teleprism.com/iccams-evaluation/LR.pdf**).** Usually, also this is in combination with an impact on learning outcomes. In Holmqvist’s (2017) review, for example, they concluded that all five models they evaluated were recommended to develop teachers’ pedagogical content knowledge (PCK) in mathematics and also thereby improve students’ mathematical knowledge.

PD teachers’ beliefs and pedagogical content knowledge

The importance of pedagogical content knowledge (PCK) to teaching—and of teachers’ beliefs (about their needs for PD)—is very well established in the mathematics education literature (Schoenfeld, 2011). The literature also recognises that it can lead to ‘deep’ change in professional practice. Hodgen, Foster, Marks and Brown (2018) report that for the teaching of mathematics, teacher knowledge (specifically PCK) is the key factor in realising the full potential of curriculum resources. The facilitation of student learning is therefore dependent upon teachers’ interpretations and transformations of subject-matter knowledge. The way to enhance this is via extended PD programmes such as the two-year ICCAMS. However, as explained previously, there is now some evidence that it is the revisiting of ideas that is important in PD rather than the duration of the programme.

PCK (Shulman, 1986, 1987) has been extensively researched in mathematics education and, indeed, remains an issue of concern in the literature—for even well-qualified mathematics teachers. Although it is increasingly accepted throughout the world that teachers’ PD is a priority, research reveals the ineffectiveness of the majority of school-initiated programmes (Mokhele and Jita, 2012) due to the dissonance between many PD sessions and teachers' personal needs and expectations. In short, research (for example, Gresalfi and Cobb, 2011) suggests that there is a willingness by teachers to change their instructional practices if they believe that the effort is worthwhile. Therefore, self-initiated programmes could be the most effective as the outcomes are more meaningful for the teachers (Mokhele and Jita, 2012). These findings are further corroborated by Ruthven et al. (2017) and we should also note the role of time to allow reflection and development of sustainable teaching practices, as mentioned earlier.

Having examined what works for effective PD in secondary mathematics and the associations of PD with teachers' dispositions and pedagogical content knowledge, an important question to ask about any intervention like ICCAMS, then, would be: which teachers' needs are or were met by the PD and in what way did the various teachers respond to the programme? We can add to this questions about the commitment of the system and the schools in relation to the depth of challenge the PD posed to the teachers' previous practice.

Challenging and changing existing professional practices

As already noted, it is usually challenging for teachers to 're-author' their professional identities to incorporate change after their participation in PD (Darragh and Radovic, 2019). A number of studies have reported findings in relation to changes in teaching practices after PD in mathematics (Chirinda and Barmby, 2017; Panizzon and Pegg, 2008; Van Zoest, Breyfogle and Ziebarth, 2002; Watson and Beswick, 2011). Studies on PD participation also reported changes in motivation (Gresalfi and Cobb, 2011), and also improvements in pupil performance (Watson and Beswick, 2011). The literature reported evidence of impact on various aspects of teacher practice, including questioning teachers' pedagogical practices, attention to cognition, and teachers becoming facilitators rather than directive teachers. A large U.S. study, for example, reported steady and statistically significant changes in teachers' mathematical discourse, instructional clarity, and the development of students' mathematical thinking, but not in student interactions or in the use of multiple representations (Copur-Gencturk and Papakonstantinou, 2016). Much PD encourages teachers to move from teacher-centred approaches to more collaborative student-centred teaching with reported successes in encouraging verbal engagement using various questioning strategies and problem-based collaborative group work (Keast, 2015; Lee, 2014; Howard and Miller, 2018).

The literature so far also suggests that FA-informed pedagogies can have an impact on learning outcomes (including metacognitive, attitudinal, and substantive attainment outcomes) but that the impact may be significantly affected by how much the pedagogy is instrumental rather than engaging learners in dialogue and reflection. The ICCAMS intervention is different from many generic FA interventions in focusing change on specific lessons, materials, tasks, and discussions that are hypothesised to help teachers implement a more effective FA pedagogy. Even if the pedagogy implemented is not strongly FA-informed generically (for example, in the quality of discussion and reflection in the classroom) it is expected that the tasks and materials will afford some reflection and learning in the specific contexts of the maths tasks in the manual.

This in turn directly links to the association of FA with teaching practices in mathematics more generically defined. Formative practice is an important part of connecting the teaching with 'what the learner already knows' and hence FA is one important part of 'connectionist' teaching, which is indeed the antithesis of 'transmissionist', teacher centred, 'delivery' pedagogy (Askew et al., 1997a; Askew et al., 1997b; Swan, 2006; Williams et al., 2008). The emphasis in transmissionist teaching is on 'delivering' to the students the knowledge and practices that are valued in the institution (for example, for tests) whether or not this is useful for the learner in the long term. It was extensively shown in literature on mathematics education that a scale that measures transmissionism, building on the work of Askew et al. (ibid.), as developed and validated for (self-report) perceived teaching of ages 11 to 20 (Pampaka, Pepin and Sikko, 2016; Pampaka and Williams, 2016; Pampaka, Williams and Hutcheson, 2012a; Pampaka, Williams, Hutcheson, Wake, Black, Davis and Hernandez - Martinez, 2012b), has been useful in explaining differing learning outcomes for students (both cognitive and emotional) and thus transmissionist teaching can have a key mediating or moderating role in the context of this study. In particular, both the validation process of such a scale, as well as its further use with student and teacher samples, provided consistent evidence of an association between teachers' self-reported transmissionism—as well as students' perceptions of transmissionist teaching—with students decreasing mathematics dispositions (ibid. and Pampaka, forthcoming). It seems likely then that such a measure might moderate or mediate the effects of the intervention on learning outcomes.

Evidence of impact on student learning, emotions, and motivations

A number of PD programmes have sought to improve student learning and emotions/motivations, both internationally and in the U.K. In fact, effective programmes for improving mathematics outcomes involving PD of teachers are widely reported (Foster et al., 2013). Findings in relation to the impact of various PD/FA programmes suggest, among others, increased students' perceived autonomy, competence, interest, and intrinsic learning motivation as well as maths performance (Kiemer, Gröschner, Pehmer and Seidel, 2015; Ostermeier, Prenzel and Duit, 2010; Pournara et al., 2015; Watson and Beswick, 2011). Others have reported differing outcomes: an RCT in Italy, for example, evaluating a PD for teachers on student maths achievement, found no significant impact on maths attainment but some effects on teaching practice and student attitudes (Argentin et al., 2014). Another RCT provided evidence that student-centred instructional practices are correlated with higher student achievement but also warned that, when implemented poorly, such curricula can do more harm than good (Ikemoto, Steele and Pane, 2016).

In sum, these studies do suggest that improvements in learning outcomes can sometimes be made but most of the time the evidence is more conclusive in relation to reported changes in teaching practices and learners' attitudes.

We focus next on particular evidence for multiplicative reasoning and algebra, which are central in the ICCAMS approach.

Evidence for multiplicative reasoning and algebra integrating teacher and learning aspects

Evidence focusing on multiplicative reasoning

The work of Streefland (1991) and Van Galen, Feijs, Figueiredo, Gravemeijer, Van Herpen and Keijzer (2008) on fractions, percentages, and proportions with Grades 4 to 6, which corresponds to the previous stage to that of ICCAMS evaluation age groups, is a relevant starting point for multiplicative reasoning. Their findings have been further corroborated more recently in the longitudinal analysis of Desimone, Smith and Phillips (2013) on the types of PD that can change teaching practice in ways that increase student achievement. They found that when the teachers focused more on advanced mathematics topics (fractions, distance problems, solving equations with one unknown, solving two equations with two unknowns, and statistics) and emphasised solving novel problems, student achievement grew quickly as compared to when they focused more on basic topics . They also found that when teachers participated in PD that focused on mathematics content or instructional strategies, they were more likely to teach in ways associated with student achievement growth. This resonated with a study that indicated that teachers may need support to develop sufficient subject knowledge and pedagogical content knowledge to teach proportional reasoning, which may be more challenging for many students as it involves multiplicative thinking and making multiple comparisons (Hilton and Hilton, 2019).

In the U.K., the Improvement in STEM Education (epiSTEMe)[13] project (2008 to 2013) used a randomised field trial looking at four subject topics in early secondary school: forces, fractions, probability, and electricity. Ruthven, Mercer, Taber, Guardia, Hofmann, Ilie, Luthman and Riga (2017) reported on the field trial of this classroom intervention with the associated PD, and dialogic teaching as its distinctive feature. Even though observations in the intervention schools showed that in one of the modules there was more dialogic teaching compared to the other modules, the evaluation results from the first implementation of the intervention found that learning gains, including in mathematics, were not greater for the epiSTEMe group compared to the control groups. There was neither any difference between the two groups in the opinion of the students about their classroom experience or in attitude changes towards physical science and mathematics.

In 2013/2014, the Department for Education focused on the teaching of proportional and fractional relationships in mathematics (Boylan, Demack, Willis, Stevens, Adams and Verrier, 2015). Their Multiplicative Reasoning Project (MRP) encouraged 60 teachers in 30 schools across three regional NCETM networks to make use of mathematical models, visual approaches, and problem solving strategies to support learning in these mathematical areas. Although the work was led by PD leaders and supported by university researchers, there were no statistically significant benefits for

[13] Final report: https://www.educ.cam.ac.uk/research/projects/episteme/epiSTEMeFinalReport.pdf

learners' general mathematics achievement within the seven-month timeframe of the project. This outcome was felt to be due to the length of the project and that a targeted, longer-duration intervention would be more successful, for example, such as ICCAMS.

Evidence focusing on algebra

A meta-analysis (Haas, 2005) of various teaching techniques for algebra within 35 studies from 1980 to 2002 identified six teaching methods which can have positive effects on student algebra achievement in secondary schools: (1) cooperative learning, (2) communication and study skills, (3) technology instruction, (4) problem based learning, (5) manipulatives, models and multiple representations, and (6) direct instruction. Using these categories and the outcome measures reported in the analysed studies, a concluding focus was on the question, 'Which teaching methods will produce the strongest influence on algebra student achievement?' (p.39). According to this meta-analysis, 'direct instruction' methods had the largest mean effect size (0.55, with 21% percentile gain for the experimental group students), followed by problem based learning (mean effect size 0.52 and 20% percentile gain), and then manipulative models and multiple representations (mean effect size 0.38 with 15% percentile gain). The remaining three categories of teaching techniques were also considered influential, albeit to a lesser degree. What is interesting from this study is perhaps the definition of direct instruction as 'establishing a direction and rationale for learning by relating new concepts to previous learning, leading students through a specified sequence of instructions based on predetermined steps that introduce and reinforce a concept, and providing students with practice and feedback relative to how well they are doing' (p.28). Later, this category was described as similar to Marzano, Pickering and Pollock's (2001) 'setting objectives and providing feedback' category' (Haas, 2005, p.31). We note that these definitions include formative assessment aspects tied closely to the concepts to be learnt. This resonates well with the ICCAMS FA approach as well as the use of multiple representations that are central to the material developed.

Two studies focused on PD in the context of algebra and meaningful mathematics discourse (Piccolo, Harbaugh, Carter, Capraro and Capraro, 2008) and lesson planning for dialogic instruction (Heyd-Metzuyanim, Munter and Greeno, 2018). Both suggest that PD is particularly valuable in the teaching of algebra by providing teachers with more subject knowledge and consequently more confidence to 'allow' their students to ask questions.

In sum, the focus of ICCAMS on algebra and multiplicative reasoning is likely to be challenging to teachers' mathematics knowledge and the focus on problem solving is likely to require support for their pedagogical content knowledge: therefore, for both algebra and multiplication, professional development can (and should aim to) enhance teachers' mathematical knowledge but also develop pedagogical knowledge regarding (research-informed) ways to teach the two topics.

Concluding the literature review

To sum up the literature review, then, there is good evidence that formative assessment practices, soundly conceived, can impact learning outcomes, but it is a challenging pedagogy and PD needs to be carefully planned and resourced accordingly. We have pointed to the difficulties and challenges that such PD faces and will return to these issues in discussion of the ICCAMS evaluation at the end of this report. The major question the review raises for ICCAMS then is the degree to which this approach demands changes in pedagogic practice (and associated professionals' beliefs and understandings) in order to effect the learning outcomes intended, and what resources of time and professional commitment this will demand.

One can read in the literature that deep changes might take place over periods of three to five years, while short term PD programmes often do not produce concurrent improvements in learning outcomes. Successes in small scale projects may be associated with intense engagement by small numbers of teachers together with leaders of PD who have expertise in research and development. The use of well-crafted resources and tasks may be a necessary condition for teachers to start to believe in a new approach, but there is also the need for reflection on experiences with peers in networked development groups.

Before we turn to the particular details of the evaluation a note on the development/piloting stage of the current study and its evaluation follows.

Piloting and development stage

The newly developed training model for PD leads, the teacher PD, and the new resources were subject to a piloting phase with eight to ten schools in the East Midlands between January 2016 and July 2016. Data was collected by the developer and delivery teams through various sources:

- interviews with students in small groups about mathematics;
- evidence from students' work (via examples);
- interviews with teachers and students about the ICCAMS2 intervention, and
- observations of lessons and PD sessions.

Our evaluation team collected information during the end of this pilot stage (summer of 2016) from two schools in order to pilot the instruments and measures to be used in the main study and to trial interviews. Appendix 2 presents an overview of this pilot study for validating the instruments to be used in the main trial whereas the final measures are presented at later sections.

Evaluation objectives

The evaluation was designed considering the following principal question:

> Does the ICCAMS-trained teaching practice improve students' learning outcomes in Year 8, as compared to 'business as usual' teaching practice?

In addition, the evaluation aimed to investigate, and where possible to measure, the effects of the ICCAMS maths intervention on (1) changes in teaching or pedagogy, (2) changes in students' algebra and multiplicative strategies and reasoning, and (3) changes in students' disposition towards mathematics.

The following research questions (RQs) guided the impact evaluation:

Primary question

RQ1: Do students in schools implementing ICCAMS maths over a two-year period demonstrate improvements in overall mathematical attainment compared to students attending control schools?

Secondary questions

RQ2: Do students in schools implementing ICCAMS maths over a two-year period demonstrate improvements in attainment in algebra (2a) and multiplicative reasoning (2b) compared to students attending control schools?

RQ3: Are effects on attainment different for students eligible for FSM? If so, how?

RQ4: Do students in schools implementing ICCAMS maths over a two-year period change their dispositions to learning mathematics compared to students attending control schools?

RQ5: Is there an interaction between fidelity and attainment change for the treatment schools?

Implementation and process evaluation questions

The main questions to be addressed through the process evaluation are listed below:

RQ6: How, and to what extent, do the involved stakeholders (for example, PD leads, schools, and teachers) practise and adhere to the principles, guidance, and materials?

RQ6a: How much of the training have teachers attended, and how was it delivered?

RQ6b: How frequently do teachers report they implement the ICCAMS materials and for how long?

RQ6c: To what extent do ICCAMS materials and PD support PD leads to deliver ICCAMS? Are there ways in which these can be improved?

RQ7: How, and to what extent, does the method by which training is offered (for example, PD lead or cascade) relate to how ICCAMS is delivered in the classroom?

RQ7a: Are there differences in fidelity between lead and cascade teachers?

RQ7b: What are the contextual factors that afford or constrain the quality of implementation and the cascading in school?

RQ7c: To what extent do ICCAMS materials and PD support the cascade model of training? Are there ways in which these can be improved?

RQ8: How do students engage with ICCAMS, including lessons, materials, and related practices?

RQ9: What relevant mathematics and PD systems and practices are in place in schools randomly allocated to the 'business as usual' control group?' And how do these relate with the impact seen on primary and secondary outcomes (that is, attainment and dispositions) identified at classroom or school level?

RQ10: To what extent do pedagogical factors, (for example, transmissionist or connectionist approaches, confidence in teaching ICCAMS, and fidelity of the intervention) mediate or moderate the impact of ICCAMS on primary and secondary outcomes? (The measurement of these factors will be detailed later.)

The links to the latest version of the protocol (amended November 2018, from initial version of August 2016) and the Statistical Analysis Plan (SAP) are provided below:

https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/ICCAMS_protocol_updated_June_2018.pdf
https://educationendowmentfoundation.org.uk/public/files/Projects/ICCAMS_SAP_.pdf

Ethics and trial registration

Each of the participating institutions on the ICCAMS project and its evaluation has received ethics clearance within their institution. Ethical approval for the pilot stage of the independent evaluation was granted by the University of Manchester Research Ethics Committee 6 on 14 June 2016 (ref: 16348) and ethical approval for the main trial was granted by the University of Manchester Research Ethics Committee 1 on 9 September 2016 (ref: 16405). Ethical approval for Phase 1 of the ICCAMS2 study was granted by the University of Nottingham's School of Education Ethics Committee on 8 October 2015 (ref: 2015/938/MO). The application for both Phase 1 and Phase 2 parts of the ICCAMS2 project was granted by Durham University's School of Education Ethics Sub-Committee on 11 December 2015 (ref: 2245).

Parental opt-out consent was sought and obtained for collecting and using data for the trial. Opt-in consent was sought from students and teachers for observations, interviews, and surveys. At the beginning of the ICCAMS project in 2016, all the schools which had agreed to be involved signed a school participation agreement (see Appendix 3). Subsequently the parents of Year 7 students in each of these schools were given information about the project which allowed them to opt their child out of the data sharing and processing for this research project at the beginning of the 2016/2017 academic year (see Appendix 4). This process was repeated when their child entered Year 8 at the beginning of the 2017/2018 academic year.

The study was registered at http://www.isrctn.com/ISRCTN12649501

Trial registration number: ISRCTN12649501

Data protection

Data for this project has been collected and used in line with public interest (Article 6 (1)(e) of the General Data Protection Regulation) to carry out research and inform future educational provision in relation to mathematics teaching and learning. As the data collection of this evaluation was split before and after the introduction of the new General Data Protection Regulation (GDPR) and the Data Protection Act 2018, the relevant stakeholders (schools, teachers, and parents) were informed accordingly about the legal basis of processing such data (see letters and links to institutional privacy notices in Appendix 5). These letters along with the participant information sheet attached with the test and the questionnaire for the students (Appendix 7A) detailed our approach to demonstrating GDPR compliance, including how we are protecting individual data subjects' rights, including their right for withdrawal and the way to do so, the purposes for data processing, and the way we were to share data between the involved teams. In particular, parents and students were provided with an opportunity to withdraw from any data processing as part of the research to ensure that they have no objection to their data being processed in this way. This was in order to demonstrate that the processing does not impinge on anyone's rights and to meet our responsibilities under the BERA Ethical Guidelines for Educational Research (particularly regarding informed consent, openness, and disclosure). Each institution was accorded data controller status for processing the data for their own, separate research purposes and there was a data sharing agreement in place for the transfer of such data between the teams.

Project team

The various roles and responsibilities of the teams involved as noted below are also described in the protocol and can also be seen in Table 17.

Development and delivery teams

The intervention was developed by a team initially based at Nottingham University: they have been responsible for developing and piloting the ICCAMS programme and the professional development.

Name	Role
Jeremy Hodgen	Principal developer of the intervention (moved to UCL Institute of Education on 1 September 2017).
Dietmar Küchemann	Support and guidance to PI and developer of new lessons.
Marc North	Responsible for developing pilot PD and PD sessions (until 31 August 2016).
Colin Foster	Additional support and guidance on PD and lesson materials (from 1 September 2016; moved to University of Leicester 1 January 2018).
Kanchana Minson	Project administrator.

The Durham University (Delivery) Team was responsible for recruitment, delivery of the intervention, and monitoring fidelity:

Name	Role
Vic Menzies	Principal investigator at Durham University responsible for trial recruitment, PD lead recruitment and training, and delivery of the scaled up ICCAMS model to schools. (Maternity leave October 2016 to January 2018.)
Gemma Stone	Maternity cover for Vic Menzies (October 2016 to January 2018) responsible for the delivery of the scaled up ICCAMS model to schools.
Stephanie Raine	Research assistant (January 2016 to September 2016).
Jessica Hugill	Research assistant (from May 2017).
Clare Collyer	Research administrator (until October 2017).
Mary Nezzo-Thompson	Research administrator (from March 2018).
Rob Coe	Professor of Education and Director of CEM—advice on intervention and trial conduct.
Andy Wiggins	Associate director at CEM—support and advice on intervention delivery and trial conduct.

The following Maths Hubs were involved in this study:

- East Anglia
- East Midlands
- London
- South West
- Yorkshire

The PD leads were engaged directly with the development and delivery team in several day long sessions in which they were invited to work over the materials and collaborate in developing proposed plans for regional PD sessions they were to teach.

Evaluation team

The Manchester University (evaluator) team was responsible for the independent evaluation, the results of which are presented in this report.

Name	Role
Maria Pampaka	Principal investigator of the independent evaluation, lead on the statistical analysis and report writing.
Julian Williams	Professor of Mathematics Education, co-investigator and lead of the IPE.
Lawrence Wo	Research associate.
Graeme Hutcheson	Statistician and advisor.
Abate Kenna	Researcher (until August 2017), IPE
David Swanson	Researcher (from October 2017), IPE.
Jack Quinn	Researcher (from October 2017), project administration and IPE.
Diane Harris	Researcher (from 2018), literature review and IPE.

The evaluation team also employed a number of casual staff for testing, marking, and data entry.

Methods

Trial design

The main trial was designed as a cluster randomised controlled trial with two arms, as summarised in Table 7.

Table 7: Trial design

Trial design, including number of arms		Two-arm, cluster randomised controlled trial
Unit of randomisation		School
Stratification variable(s) (if applicable)		Geographic area (region) GCSE % and FSM % within each region (blocks)
Primary outcome	Variable	Mathematics attainment
	Measure (instrument, scale, source)	An amended version of MALT 13
Secondary outcome(s)	Variable(s)	Multiplication attainment Algebra attainment Mathematics disposition
	Measure(s) (instrument, scale, source)	The amended version of MALT 13 Bespoke survey instrument for mathematics disposition
Baseline for primary outcome	Variable	Maths attainment
	Measure (instrument, scale, source)	KS2 mathematics score (scaled); available from NPD
Baseline for secondary outcome(s)	Variable	Mathematics disposition
	Measure (instrument, scale, source)	Bespoke survey instrument

The main trial was designed as a cluster randomised controlled effectiveness trial with two arms. Randomisation was at school level and took place in July 2016, after all involved schools were recruited (and returned the participation agreement). The trial ran for two academic years starting in September 2016. Given this (school-level randomisation) design, two-level models were used for the key findings, however, explanatory models further took into account the clustering of students within classes (and their mathematics teachers) (see statistical analysis section).

Schools allocated to the intervention arm of the trial were trained and supported to implement the ICCAMS intervention over a two-year period. These schools were expected to run the ICCAMS intervention for all students in Year 7 initially and for the same students again when they were in Year 8 (September 2016 to July 2018). The main primary outcome is mathematics attainment measured through a test, and there were also secondary outcomes related to attainment subscales for algebra and multiplication and attitudes to mathematics (these are detailed in a later section).

Schools allocated to the control arm of the trial were encouraged to continue practice as usual during the same period. £1,500 was offered to control schools as compensation for time and to avoid attrition: £500 at the beginning of the trial (upon provision of UPNs and completion of pre-survey) and £1,000 following the completion of post-test measures. Schools in both arms of the trial were required to sign a participation agreement (see Appendix 3) before randomisation committing them to comply with the evaluation protocol whichever arm they were allocated to. It should be noted that the term ‘control’ is used here as a convention and without the strict connotation of ‘controls’ as in the experimental designs. It is rather used interchangeably with comparison or comparator group and the complexities around this and ‘business as usual’ will be discussed at later sections.

During initial discussions with the EEF and the developers of the intervention we had considered the possibility of implementing a factorial design or alternative three- and four-arm designs. The factorial design scenario involved intensity of PD (that is, low vs high) and materials used (ICCAMS vs other) as the two factors: that would have involved a 'light ICCAMS three-day PD' and 'full ICCAMS nine-day PD' and a lesson study with low and high intensity involving an alternative intervention (such as CAME or Maths Mastery). The possibility of including a control group along this factorial specification was also considered, which would have made the design an incomplete factorial. Alternatively, if the project was to evaluate whether the material *or* the PD have an effect then a three-arm (control, ICCAMS materials without PD, ICCAMS materials with PD) or four-arm design (as three-arm with another treatment ICCAMS material with light PD) would be ideal. At one of the initiating meetings of the evaluation, though, (23 February 2015) it was agreed that the Maths Mastery and the ICCAMS-light options were insufficiently well-developed to use as the basis for either a factorial or a third (or fourth) arm. It was, thus, subsequently agreed to proceed with the chosen design two-arm (business as usual) design, which is considered advantageous in terms of the balance between scientific rigour, ethical considerations, and goodness-of-fit with the study aims and hypotheses. There have not been any significant changes to this design since the main trial commenced in September 2016.

Participant selection

According to the agreed protocol, eligible schools were mainstream English state secondary schools or middle schools with more than two class intakes for Year 7 (ideally not in special measures) and with, ideally,[14] higher than average levels of FSM eligibility. Schools were only eligible to take part in the study if they agreed to all of the study requirements outlined in the Participation Agreement between the universities and schools and the form was signed by the headteacher (see Appendix 3). A cautionary note is needed here: despite the requirements for data provision, at later stages of the project and its evaluation some schools were not fully compliant to these requirements (that is, some schools did not provide all the data later and some schools did not provide any data at all), however, they were not considered as not eligible at the recruitment or randomisation stage (as this had to take place before any further data collection).

The trial schools were recruited by the delivery team, supported by NCETM and the Maths Hubs, and Nottingham (development team) in five regionally-based groups to facilitate the hub-based PD. Discussions with NCETM and the EEF regarding further recruitment criteria concluded that other related projects (for example, the KS3 Multiplicative Reasoning Project) were considered not similar enough to prevent schools in the KS3 Multiplicative Reasoning project from taking part in the ICCAMS project (as was initially decided at the first version of the protocol).[15] As per the protocol, recruitment had to only monitor schools also taking part in the Schools, Students and Teachers Network SSAT trial, 'Whole school Embedding Formative Assessment Project', or other programme which was deemed to be too closely related. In order to ensure minimum recruitment of these schools, the EEF provided a list of schools taking part in the above programme during this stage of the study.

The initial aim was to recruit between 100 and 110 schools (to ensure that the trial was sufficiently powered, with a Minimum Detectable Effect Size (MDES) of at least 0.15,[16] with a low level of school-level attrition) to take part in this trial (55 in each arm). All eligible schools within each area were invited (in writing) to the local recruitment events, or to attend a webinar, as well as being sent information about the project. The developer and delivery teams worked with the Maths Hubs in each area to increase the prominence and reach of the project to support recruitment. The Maths Hubs were recruited by the delivery team and NCETM. All the Maths Hubs were invited to apply: Maths Hubs were offered £3,000 towards costs for their involvement in the project. Maths Hubs were asked to support recruitment and to provide cover for a nominated Maths Hub ICCAMS lead to attend all the training sessions. Of the 13 Maths Hubs that submitted Expressions of Interest, Maths Hubs were paired if it made geographical sense so that the five areas were made of four Maths Hub pairs and one individual Maths Hub. Decisions on selection of Maths Hubs and areas were

[14] It should be noted that this is how recruitment criteria were agreed on the protocol: upon final recruitment of schools for randomisation it was observed that none of the 109 schools was on special measures. Further information on other demographics will be presented in the relevant sections later in the report (Randomisation and Comparisons at Baseline).
[15] This restriction was part of eligibility criteria defined at the first version of the protocol (https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Protocols/EEF_Project_Protocol_ICCAMS_University_of_Manchester_March_2016.pdf), which was revised for the second version.
[16] A series of simulations were also run with MDES 0.1 and 0.2

made based on the possibility of Maths Hub pairings, proximity to PD leads, involvement of Hubs in the EEF Year 2 Mathematical Reasoning Trial (minimising Hubs involved in both), and the number of other projects the Maths Hub was already committed to. When Maths Hubs were paired, the pairing received the £3,000 rather than both the individual Maths Hubs.

The Maths Hubs sent out information about the project to schools registered with them and promoted the local events. PD leads also supported recruitment of schools to the project through using their local network contacts and by attending local recruitment events. At a national level, NCETM promoted the project by including it in their news updates to schools and in their termly magazine. The Tes also included a short article to raise awareness of the project. The project recruited 109 schools to the trial that met the criteria and completed the participation agreement, spreading across the five areas as shown in Table 8 (names of Maths Hubs are removed to ensure anonymity and regional codes as listed below are to be used for the presentation of results).

Table 8: Final recruitment information (school-level) per region

Maths Hub centre	Number of schools recruited	Region
1	19	Region 4
2	20	Region 3
3	19	Region 1
4	25	Region 2
5	26	Region 5

All students in Year 7 at the beginning of the 2016/2017 school year were the target cohort (excluding those without parental consent) with an estimated sample of 11,000 students at the beginning, based on an estimated average of 100 per school. Since the target year was Year 7, it was not possible to collect Unique Pupil Numbers (UPNs) from the schools or obtain opt-out consent until September, which was after randomisation.

Children eligible for FSM were a subgroup for this trial; the effect of the intervention was analysed within this subgroup in view of the EEF's primary remit of narrowing the attainment gap for such students and in line with differential gains established for children from poorer socioeconomic backgrounds for related universal programmes.

The corresponding roles and responsibilities were as follows:

- The delivery team (with support from the developer team) was responsible for recruiting nine Maths Hubs across five centres.
- The evaluation team randomised schools within each hub to intervention or 'business as usual' control in July 2016.
- The delivery team was responsible for the delivery of the intervention and maintaining contact with intervention and control schools during the project (with guidance from developer team).
- The evaluation team was responsible for conducting the primary and secondary outcome assessments collected at the end of the second academic year as well as the baseline survey measures collected in September 2016 (see details next).

Outcome measures

The primary outcome for this evaluation is mathematics attainment with secondary outcomes covering two subscales of attainment focusing on multiplication and algebra accordingly, along with an attitudinal measure capturing students' mathematics dispositions. The details of the measures used for each of these outcomes at the different stages of the evaluation are provided below for primary and secondary outcomes and baseline measures.

Primary outcome

The primary attainment outcome for this evaluation is the raw (and Rasch logit, see more details later) scores on a slightly modified version of the Mathematics Assessment for Learning and Teaching (MALT) test for Year 8 (MALT 13). Revisions included the removal of two items and the addition of four algebra related items to strengthen this dimension (the measures have been piloted and results were reviewed by independent members of the EEF's Evaluation Advisory Group). This assessment is a test of general mathematics attainment but also includes some conceptual elements of mathematics. The original MALT test was complemented with the extra items to strengthen the secondary outcome measures of interest in this evaluation, and in particular algebra, as detailed below. Attainment was measured at the end of Year 8 with this revised MALT test (the paper test took 45 minutes). The test included 41 questions, five of which had two marked parts resulting in a maximum total score of 46 points.

Administration of the tests was conducted by our team from the University of Manchester and implemented under exam conditions in schools between 11 June and 13 July 2018 with our invigilators and markers blind to condition. In particular, test papers were sent by Manchester to the exams officer in each school to arrive before the test date. The papers were kept securely at school until the test date, which was organised at each school's convenience in the given window.

Test administrators from our team were present to administer the test and check test conditions. We briefed any invigilators and staff on the procedures for the test. We sent two people to each school in almost all cases (unless the school was very small or very large, and we adjusted numbers accordingly). Our test administrators took the papers from the school afterwards. Details for the guidance provided to invigilator and schools are presented in Appendix 8.

Reports from our testers and invigilators provided evidence that the schools were extremely supportive of the testing arrangements. Staff and freelance invigilators were made available for the tests in addition to the University of Manchester's team. Members of the school's senior leadership team and the teachers themselves spoke about how this was the first external examination the students had taken at secondary school and that it was useful for them to experience exam conditions. In practice, there was the usual nervousness amongst the students that could be expected when sitting an external examination. In the majority of cases, the staff knew who might be especially intimidated by sitting a test in a large assembly hall or gymnasium and took preventative measures to minimise disruption to others.

The tests were marked by research assistants and postgraduate students with experience of marking maths tests. The markers were blind to condition. They marked the tests and then entered the data into Excel files. This took place between July and September 2018. A third of scripts were checked for consistency. Please see instructions for markers in Appendix 9.

The resulting data from the tests was recalibrated as there had been some further amendments since the pilot; more details of this procedure are provided under the measure construction and validation section of the statistical analysis, along with the psychometric overview of the resulting measures (for primary and secondary outcomes as well as the other measures constructed for the IPE and explanatory models). See also Appendices 11 and 12 with Rasch measurement results.

Secondary outcomes

For the impact evaluation the following three secondary outcomes were also used:

- an attainment subscale of MALT of 'multiplicative reasoning' (maximum of 20 points);
- an attainment subscale of MALT of 'algebra' (maximum of 11 points: and
- students' dispositions towards mathematics.

The attainment subscales were defined at the pilot stage and were further developed with the calibration of the measures at post-test. The distributions of these measures are shown before the description of the primary analysis approach.

Students' disposition towards mathematics was measured using the nine items presented in Figure 8. This measure had been validated as part of our recent ESRC work with secondary students and drawing on previous versions for older students (Pampaka, Williams, Hutcheson, Black, Davis , Hernandez-Martinez and Wake, 2013; Pampaka and Wo, 2014). In this work and in the pilot study of ICCAMS, which was distributed to both Year 7 and Year 8 students, we administered 18 items about students' feelings towards mathematics and found that such items fall into two related attitudinal measures: we identified these as 'maths disposition' and 'maths self-identification'. Previous work has also found the former to be more sensitive to teaching practices (Pampaka et al., 2012b; Pampaka and Williams, 2016; Pampaka, Pepin and Sikko, 2015). Therefore, this measure was considered more useful to detect potential changes due to teaching and learning practice changes that are expected with the ICCAMS intervention and thus we chose this measure using just the relevant nine items as a secondary outcome for this evaluation instead of the self-identification subscale, which is thought to be harder to change in the shorter-term.

Student dispositions were collected at the start of 2016/2017 academic year as an additional baseline measure and at the end of 2017/2018 along with the primary attainment outcome. The first data collection took place via hard copies of surveys posted to schools in September 2017 by the evaluation team along with a free-post address to return. The second survey was attached to the student test and was administered alongside the MALT tests as explained earlier (see Appendix 7).

As presented in the protocol and SAP, the plan was for the primary and secondary attainment outcome measures to be first analysed using raw scores. However, owing to the small number of items in the secondary outcomes (MALT subscales) and the ordinal nature of items for the attitudinal measures for students, these outcomes were also calibrated using the Rasch modelling framework (Bond and Fox, 2007; Wolfe and Smith Jr., 2007a, b), which allows for objective measurement (see details later). Models were thus run for both raw and Rasch (logit scores); see details under Statistical Analysis section.

It should be noted that the questionnaires included more items aiming to capture other elements of teaching practice and students' perceptions of it: these are detailed under the Measurement and IPE section.

Baseline measures

Baseline measures of attainment were obtained from the National Pupil Database (NPD). These are the scaled scores in mathematics (KS2_MATSCORE)[17] collected when pupils were in Year 6. Schools were asked to provide name, unique pupil number (UPN), gender, the EverFSM 6 status, date of birth (DOB), KS2 maths result, class, and maths teacher for all eligible students at baseline. This enabled us to collect KS2 results from the NPD and match pupils to their teachers. When schools did not provide UPNs, matching (of students who had parental consent) was performed with information we had from surveys at baseline or the names and dates of birth provided by students during the testing stage.

The extent to which this happened is shown later on in the report with the descriptive and missing data analysis.

Sample size

The sample size—the number of schools in each cluster—needed for each of the two arms of this study was determined at protocol stage based on the following assumptions:

- 80% power and alpha of 0.05;
- a minimum detectable effect size of 0.15; this was deemed a worthwhile effect given the estimated cost of the intervention and the cascade delivery of PD within schools; this was further supported by a previous evaluation of ICCAMS (Hodgen et al., 2014) which also suggested an effect size of this order is a reasonable target;
- ICC of 0.12 based on a combined consideration of suggestions and assumptions in relevant literature[18] (Hedges and Hedberg, 2007; Spybrook and Raudenbush, 2009); and
- a pre-test post-test design with 0.65 correlation.[19]

With these assumptions, it was estimated with PASS software (Donner and Klar, 1996; Donner and Klar, 2000) that a minimum of 50 schools would be required per trial arm (assuming the number of students in Years 7 and 8 in these schools ranging from 75 to 150 based on the discussions at that time and noted on the protocol). A target of 110 schools was thus set for recruitment and an estimated sample size of 10,000 students. The trial eventually recruited 109 schools that were then randomised (see next section). The average size of recruited schools was larger than initially estimated (about 200 students per school) with potential implications for the resulting power. The MDES at the various stages

[17] As opposed to the raw scores, which are not available in the most recent NPD tables, https://www.gov.uk/guidance/scaled-scores-at-key-stage-2

[18] A cautionary note should be made here in relation to not considering the class or teacher level in such calculations as recently suggested by Demack (2019) for the possibility of a three-level design being more appropriate for mathematics. We return to this later as we have also run three-level models for certain outcomes.

[19] This was informed by a combination of references and guidelines and with a conservative decision in mind to ensure the required MDES. https://ies.ed.gov/ncee/pubs/20114033/pdf/20114033.pdf, https://v1.educationendowmentfoundation.org.uk/uploads/pdf/Pre-testing_paper.pdf

(protocol, randomisation, and analysis, as appropriate) was calculated using a tool (Troncoso, 2020)[20] based on the EEF guidelines for two-level clustered randomised controlled trials. The results are reported in Table 17.

Randomisation

Random allocation was performed at the school level in July 2016 (with baseline measurement planned for September to October 2016) after the receipt of the school file with relevant school-level information (%FSM and %GCSE A* to C). Randomisation was planned to be performed within each regional hub with an expected maximum of 30 schools each, and thus using block stratified randomisation (Torgerson and Torgerson, 2008). In order to ensure balance in regards to previous attainment and proportion of FSM students, blocks were expected to be defined by the proportion of students in each school to achieve five A* to C GCSEs in the 2015 examinations (above median and below median) and the proportion of FSM students in each school (above and below median). This implied that there were up to four blocking variables, or strata, made up of the combinations of these two variables. Preliminary investigation of the given school information based on the medians of the two strata (FSM and GCSE %) within each area revealed some problems, especially with confounding of the two variables in some areas. It was observed, for instance, that high FSM schools with medium GCSE were concentrated in one area, and another area did not have any schools of high FSM%. In order to account for these patterns, deal with the missing information (information not available)[21] for some schools, as well as ensure balance in the overall design and school split, it was considered more useful to define the groups/blocks based on three categories per stratum as shown in Figure 5 and with the steps detailed in Appendix 10 (which also includes the blocks by regions).

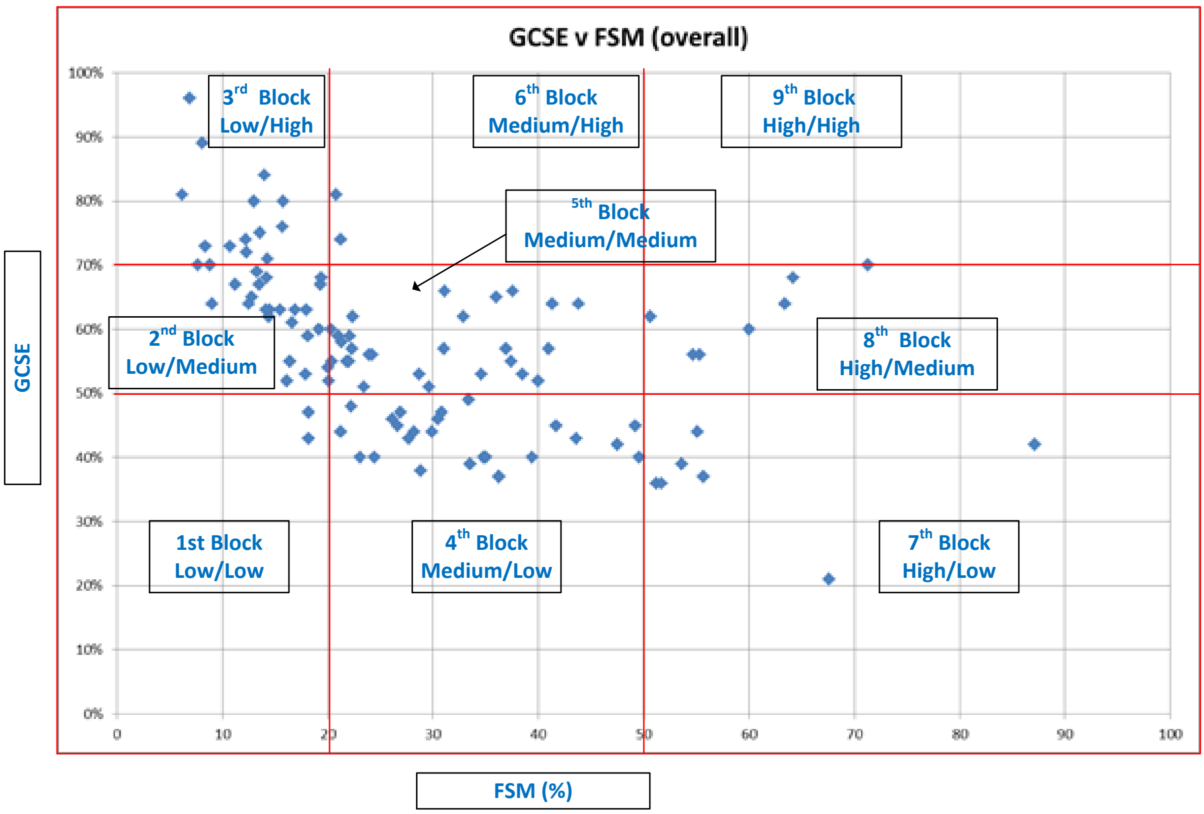


Figure 5: Block definition for randomisation within regions

As a result of this process, there were 55 schools assigned to the experimental group and 54 to the control group in total. The distributions of FSM and GCSE, according to the information provided at that time (that is, from the list for randomisation)[22] are as shown in Table 9.

[20] https://patricio-troncoso.shinyapps.io/mdesapp/

[21] From the 109 schools in the list we received for randomisation there were two new schools without recorded FSM or GCSE result data and seven additional without GCSE, which were either recently opened or previously or currently middle schools.
[22] These descriptors will be revisited in the Results section with updated information and matched when possible with the NPD.

The detail of the randomisation process was recorded (the anonymised school list along with the algorithm can be seen in Appendix 10) and the outcome was shared with the delivery team. The schools were then informed (by the delivery team) of their random allocation in July in order to make the necessary arrangements needed for the teachers to attend the ICCAMS PD sessions.

Table 9: Average percentage of FSM and GCSE of allocated schools by arm and area (n, m: number of schools with available information, number of schools with missing information)

	Hub Area [Region]					Whole Sample
	Hub 1 [4]	Hub 2 [3]	Hub 3 [1]	Hub 4 [2]	Hub 5 [5]	
Average of FSM	**(19,0)**	**(20,0)**	**(18,1)**	**(24,1)**	**(26,0)**	
Control	26.52 (9)	39.15 (10)	24.02 (9)	30.21 (13)	21.62 (13)	28.15
Intervention	31.51 (10)	38.47 (10)	21.39 (9,1)	29.65 (11,1)	23.77 (13)	28.82
Total	**29.15**	**38.81**	**22.71**	**29.95**	**22.7**	**28.48**
Average of GCSE	**(17,0)**	**(19,1)**	**(18,1)**	**(22,3)**	**(24,2)**	
Control	51.57 (7)	65.11 (9,1)	55.11 (9)	47.5 (12,1)	63 (11,2)	56.4
Intervention	52.8 (10)	63.8 (10)	56.22 (9,1)	52.9 (10,2)	60.15 (13)	57.4
Total	**52.29**	**64.42**	**55.67**	**49.95**	**61.46**	**56.89**

Statistical analysis

In this section we detail the analytical approach of this evaluation, which is also noted in the Statistical Analysis Plan. We start with a list of amendments made to the SAP, we then explain the data preparation process in order to merge the multiple data sources. We summarise our approach to validation and construction of the measures of interest and then present the details of the models for primary and secondary analysis, analysis in the presence of non-compliance, and subgroup analysis.

Overview of analytical approach

The analysis was in line with the SAP apart from the following amendments (which will be detailed at appropriate sections):

- Additional models were run with a new categorical variable to denote the type of teacher at allocation (lead or cascade) as this was of central focus for the process evaluation.
- Compliance analysis was performed both with the (school-level) fidelity measure as well as an additional instrumental variable approach (which was not fully prescribed in the SAP).
- It was also considered important to model students' perceptions of teaching practices at the end of the study based on other variables included in the main outcome models as well as teacher related characteristics (this was mainly driven by ongoing findings from the process evaluation as well as to address the limitation with missing data resulting from teacher survey returns).
- Sensitivity analyses involved modelling primary and secondary outcomes with various specifications and subsamples: in particular we run the models with (1) the condition variable as reported at the end of Year 8 (as opposed to randomisation, for the intention-to-treat models), (2) with the subsample when removing the cases who changed school and those schools which dropped out of the intervention at some point during the project time, and (3) with the subsample with complete information in all variables (that is, complete case analyses).

Data preparation

Before performing the main analytical stages in response to the main research questions for both the impact and process evaluation the following steps were necessary and worth a bit more detail to help the reader with the main analysis and findings. This included the following steps in an iterative process:

- the merging of online survey export files and hand input files;
- data cleaning (individual files) and checks for errors (for example, with marking);

- data matching across data points with main reference data shared by the delivery team;
- data matching with student surveys and school-level fidelity (when possible); and
- Rasch analysis for measure calibration and construction.

The data to be merged into the quantitative dataset was collected as shown in Table 10:

Table 10: Various data sources for linkage (before submitting for NPD matching)

	Data point 1	Data point 2	Collected by
Teacher surveys	September 2016 (DP1, teacher)	July 2018 (DP2, teacher)	Evaluation team
Student survey	September/October 2016 (DP1, student)	June/July 2018 (DP2, student)	Evaluation team
Student test	N/A	June/July 2018 (DP2, student)	Evaluation team
Student lists from schools	Year 7, 2016/2017 (Y7, pupil list)	Year 8, 2017/2018 (Y8, pupil list)	Delivery team

Merging of online survey export files and hand input files
In the first year of the study (DP1), teacher surveys were conducted on hard copies whereas in DP2 almost all teacher surveys were conducted online with only one school using hard copies. Online data was exported to Excel files and data from hard copies was entered. One spreadsheet was then created to contain all teacher survey data. Some variables only applied to the second teacher survey, which was differentiated by lead and cascade teachers.

Data matching across data points with main reference data shared by the delivery team (class lists)
Student surveys and test data were matched with the student class lists from the relevant year. DP1 student surveys were matched with the Year 7 student list and DP2 student survey and test data were matched to the Year 8 lists. The data was matched using school names, student names, DOBs, gender, class names, and teacher names. Any unmatched data was transferred to the matched file (that is, there was a row for every pupil that had returned data) whether matched or not. We then brought the matched pupil lists into one file with all student data. If files could not be matched with 100% similarity, we conducted manual checks on matching. Cases were matched manually if there was any doubt about the automatic matching. There were various challenges that called for a lot of manual checking and matching: amongst the main matching difficulties were poor handwriting, incomplete information, and survey responses from students not on the pupil lists (which could have been because of difficulties with matching due to the reasons mentioned previously). As stated above, not all data could be matched to an identical student in the pupil lists; this was therefore left unmatched, but still included in the file.

Data cleaning (individual files) and checks for errors
Samples of teacher survey scripts and student surveys were checked for errors (for example, with marking). A third of tests, over 6,000, were checked. This was done by entering pupil answers[23] and comparing it to our dataset, going back to the original scripts to find if there were errors in marking, initial data entry, or secondary data entry (when entering pupil answers). These checks provided evidence of the very high quality of the marking (from the 6,000 tests, corresponding to 276,000 item responses, inconsistency between the two entries was found in 445 items, 0.16%, and even within these cases, 10% were marked correctly) and also enabled us to correct for those errors found in these large groups.

Teacher data matching with student surveys and school-level fidelity (when possible)
The final dataset contained a maximum of four possible teacher IDs per student. Most students had one teacher per year, but sometimes two. Firstly, teacher IDs were created for those teachers that completed a survey. This was done by identifying teachers on the survey responses (using names given) and then using the school pupil lists, which contained a list of class teachers. We could therefore give teachers IDs on our main student file and the same IDs on our teacher survey file. These IDs were then used to match the teacher survey data to the main (student-level) file.

[23] This was done as part of a separate error coding exercise funded from other sources.

Some teachers from the surveys could not be identified; in addition, some teachers did not complete the survey. But all teachers listed on the student file were given a unique ID. Teachers were given a special ID if they could not be matched with a survey but were on school pupil lists.

Class IDs (Year 8) were created and allocated using the student lists collected from schools; these lists contained class and teacher names. In cases where this information was deficient we used class names and teacher names from the test scripts. But these variables, too, sometimes contained missing data and on occasions were not consistent with pupil lists; however, if class and teacher names between the test data and pupil lists conflicted, we used the test data to allocate the class ID (assuming it was most likely the student had moved classes). Where we were unsure of the Class ID these were listed as missing.

Measure construction and validation

General approach

The primary and secondary outcomes detailed earlier were calibrated using the Rasch modelling, which allows for objective measurement: the outcomes are thus similar to standardised scores. This approach was further employed for the construction of the school fidelity scores as well as the other student and teacher measures of teaching processes and confidence with ICCAMS that later informed explanatory models and also the process evaluation. The methodological approach is illustrated here using as an example the mathematics disposition measure (with the technical details presented in Appendix 11): the validation of the rest of the measures is presented in Appendices 12 to 14 and more details of each of these measures will be mentioned in the relevant sections of results.

As mentioned, the test and questionnaires (completed either by the students or teachers, see Appendix 7) included items/questions usually grouped under certain themes, like attitudes to mathematics, teaching practices, or confidence with teaching ICCAMS. In this section we will provide an overview of our approach in validating these measures (with a detailed example with results to follow in the next section).

The starting point of this approach is that there is some underlying construct (or idea or concept) behind the groups of items in the questionnaires that were brought together after studying previous research literature, looking at other instruments, discussing with relevant stakeholders, and also having been tested at the pilot stage. This is similar to what teachers do when they design and administer a mathematics test to the students: they hope to measure their mathematical attainment (maybe on a particular topic, or taught syllabus). Following the protocol and SAP, we intended to measure the following constructs for the purposes of this evaluation (including both elements of impact and implementation and process evaluation):

- mathematics attainment through a modified version of MALT 13 and two subscales of mathematics attainment: (1) multiplicative reasoning, (2) algebra;
- students' attitudinal outcomes for mathematics ('mathematics disposition');
- students' perceptions of transmissionist teaching;
- teachers' (self-reported) transmissionist teaching practices;
- teachers' (self-reported) formative assessment teaching; and
- teachers' confidence with ICCAMS (intervention only—at end of the study, post-test at Data Point 2).

Given the student and teacher responses to the relevant questions, we then attempt to validate these aforementioned constructs as measures or variables: in other words, to check whether they are viable as one dimensional 'measures' and, if not, if there are other relevant and useful or interpretable dimensions. So, following Messick (1989), our validation process refers to the accumulation of evidence to support validity arguments regarding the reported measures. We then employ a psychometric analysis for this purpose, conducted within the Rasch measurement framework and following widely accepted Rasch guidelines (Wolfe and Smith Jr., 2007a, b). For the analysis reported here we use all the three main models from the Rasch family using the Winsteps software (Linacre, 2011): The Rasch rating scale model is considered the most appropriate for common Likert type item scales, whereas the Partial Credit Model or the simple dichotomous model with binary-only responses are more appropriate for the 'attainment' measures. Our decisions about the validity of the measures were informed by different statistical indices such as item-measure fit statistics, dimensionality diagnostics, category statistics, differential item functioning, and person-item maps (Bond and Fox, 2007).

Overview of the validation process with the example of mathematics dispositions measure

We illustrate here our approach via an example with one of the measures and then summarise the remaining findings along with their implications (and include the technical details in Appendices when appropriate). The following detailed example of a standard analysis concerns the construction of the measure of students' 'mathematics disposition'. As described above, the starting ingredient in the process of measure construction is the questions or items. For the case of 'mathematics dispositions' we have used the items from the previously validated instrument with secondary school students (Pampaka and Wo, 2014).

The items in Figure 6 were, thus, given to students under the main question, 'How much do you agree or disagree with the following statements?' During DP2, these appear under 'Section A: Your feelings about mathematics' before the actual test items.

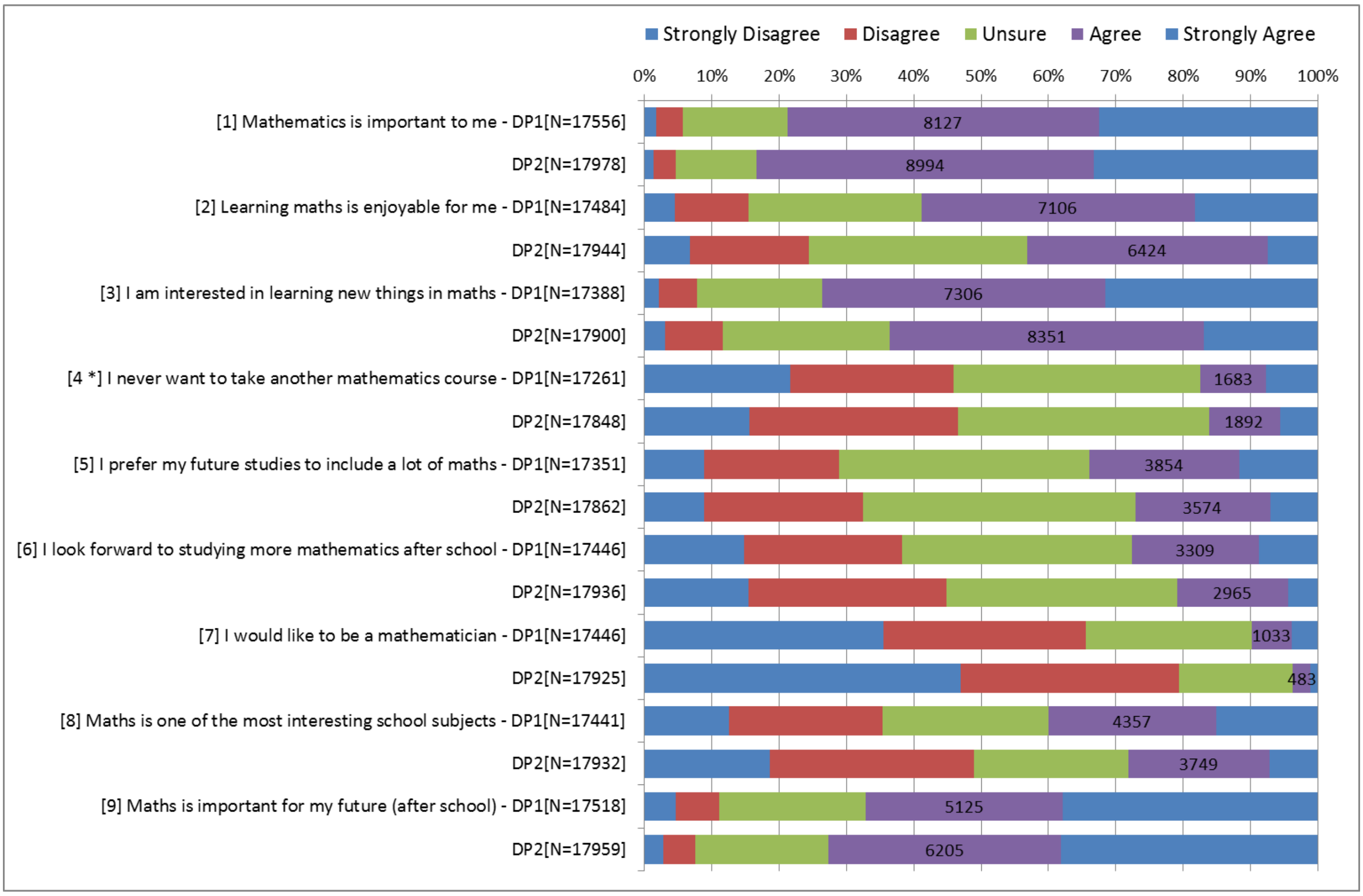


Figure 6: The items and distribution of responses for mathematics dispositions (pooled DP1 and DP2 sample)

Following our previous work, we hypothesised that these items—if appropriately scored, sometimes negatively—form together an underlying construct of students' mathematics dispositions. In order for this to be valid, however, the asterisked item [4] in Figure 6, which denotes a conceptually negative statement, needs to be reverse-scored in order for the resulting measure to be meaningful. Using, then, the tools provided by the Rasch analysis, and in particular the Rating Scale Model (because of common rating scale parameters), we create and validate this measure (as with others) based on students' responses to these items (that is, a score one to five, from 'strongly disagree' to 'strongly agree').

The ultimate outcome, if measurement is deemed valid, would be a score for each student on a logit scale (see Figure 7), which can be used for further analysis in the impact and process evaluation. Our decisions about the validity of the measures are based on various statistical indices such as item fit statistics, dimensionality diagnostics, response category statistics, and differential item functioning, with example results shown in Appendix 11. Figure 7 shows the resulting measurement scale of student scores and item 'difficulties'.

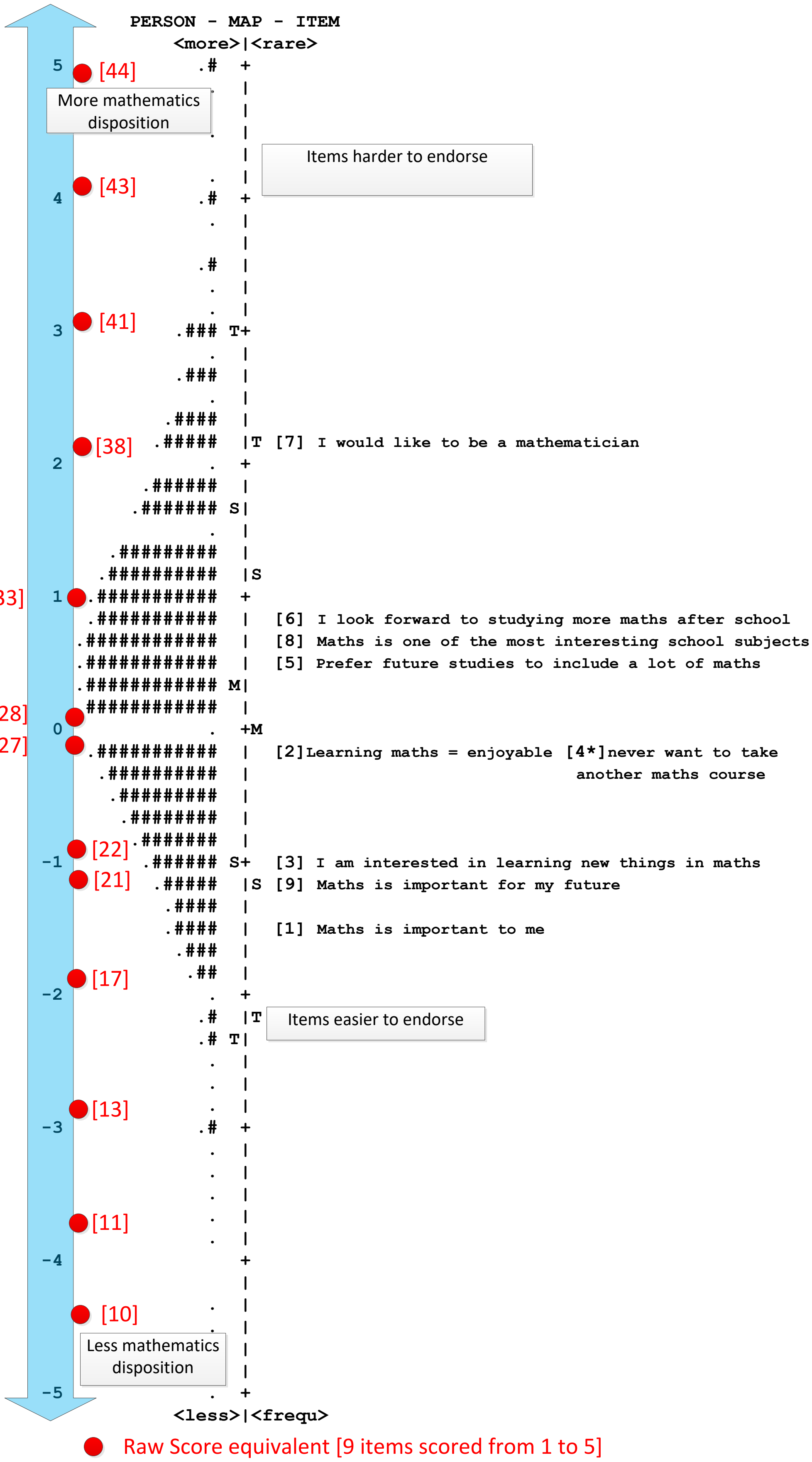


Figure 7: Item-person map and the scale for 'mathematics disposition'
(Note: '#' represents 166 students, '.' represents 1 to 165 students)

At the left end of the figure the logit scale is shown with the numbers ranging from -5 to +5 along with some examples of corresponding raw scores;[24] this is the common measurement scale for both items and persons (students). The distribution of students on this scale can be seen in the form of the histogram on the left hand side of the figure (with each # representing 166 students). The higher the place of a student in that scale the more mathematically disposed they are according to their responses to these items. On the right hand side of the students' 'histogram' the items that constitute the scale are presented, ranging from the easiest to agree with (bottom) to the most difficult. The constructed scale appears to work well for this distribution of students, with items that serve to measure students across the measurement range, and without obvious redundancy.

The same approach was employed with other measures which are to be used for both impact as well as process evaluation analytical purposes. Their distributions are summarised here. First, we report on the distribution of the other measure captured with the student surveys, that of students perceptions of transmissionist teaching (Figure 8). Three measures (two with repeated measurements) were also constructed from the teacher surveys as follows:

- transmissionist teaching at DP1: Year 7, ranging from -2.33 to +2.11 logits; and at DP2: Year 8, ranging from -3.16 to +2.51 logits;
- formative assessment practices at DP1: Year 7, ranging from -2.25 to +3.69 logits; and at DP2: Year 8, ranging from -2.49 to +4.14 logits; and
- ICCAMS confidence at DP2: ranging from -6.55 to +5.68 logits—only for intervention teachers.

The 'transmissionist teaching' and 'FA practices' measures were constructed with the items listed in Table 11 (see also questionnaires in Appendix 7, validation details in Appendix 11, and descriptors at the start of the IPE results).

Table 11: The items used to construct the measures of teachers' perceptions of transmissionist teaching and FA practice

	Item (R implies the item's scoring was reversed)	**Trans. teaching**	**FA**
1	I introduce a new topic by first determining what the students already know about it.	✓ R	✓
2	I use activities in contexts that the students can engage with.	✓ R	✓
3	I use activities which allow connections to be made between mathematical ideas.	✓ R	✓
4	I allow students to work at their own pace.	✓ R	
5	I teach the whole class at once.	✓	✓ R
6	Students start with easy questions and work up to harder questions.	✓	
7	When a student asks a question, I give clues instead of the correct answer.	✓ R	✓
8	I ask students to explain their reasoning when giving an answer.	✓ R	
9	I encourage students to discuss the mistakes they make.	✓ R	
10	Students use only the methods I taught them.	✓	
11	Students choose which questions to tackle.	✓ R	
12	Students compare different methods for doing questions.	✓ R	
13	Students work collaboratively in small groups.	✓ R	✓
14	Students discuss their ideas.	✓ R	
15	Students work collaboratively in pairs.	✓ R	✓
16	Students invent their own methods.	✓ R	✓
17	I tell students which questions to tackle.	✓	✓ R
18	I teach each topic separately.	✓	
19	I provide feedback to students on their understanding of mathematical concepts.		✓
20	I check students' understanding of maths during lessons to assess specific intended learning outcomes.		✓
21	I assess students' maths conceptions and misconceptions in order to adapt my teaching.		✓
22	I provide feedback on what students have understood in relation to what they should do next.		✓
23	I encourage students to learn from each other.		✓

[24] More information and detailed correspondence between raw scores and logit scores is presented in Appendix 11 (Table 11C, Figures 11F and 11G).

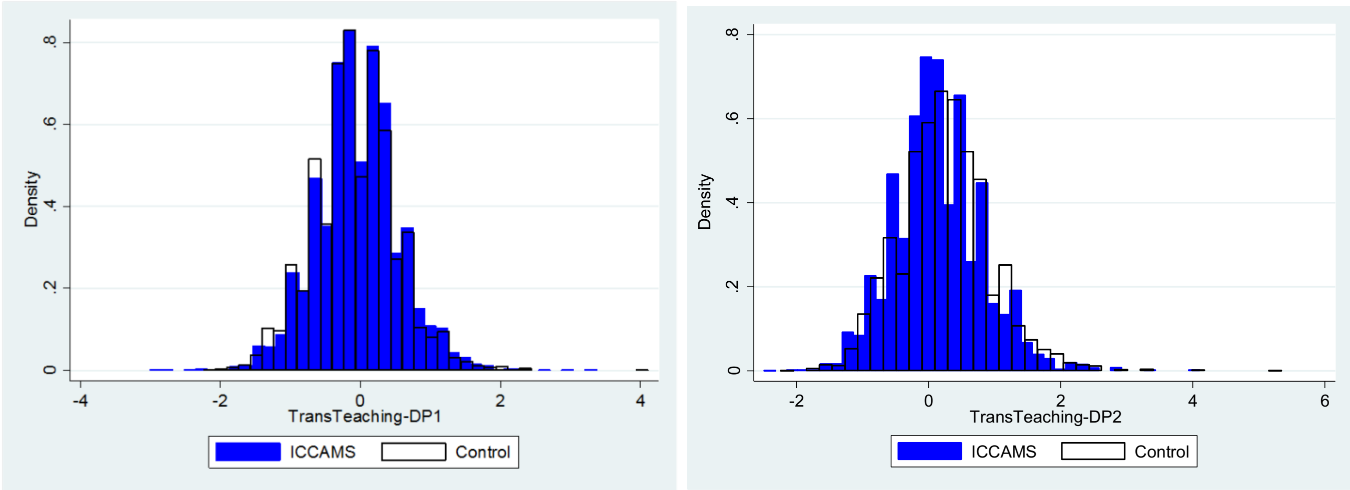


Figure 8: Distribution of students' perception of transmissionist teaching at DP1 and DP2, by condition

Please also note that the same approach was used for the construction of the measure of school fidelity, which is presented in a later section (under Methods for IPE).

Primary analysis

Before detailing the analytical approach, we present the primary and secondary outcomes to illustrate their distribution (Figure 9; see also Appendix 16 for comparative presentations between raw and Rasch scores).

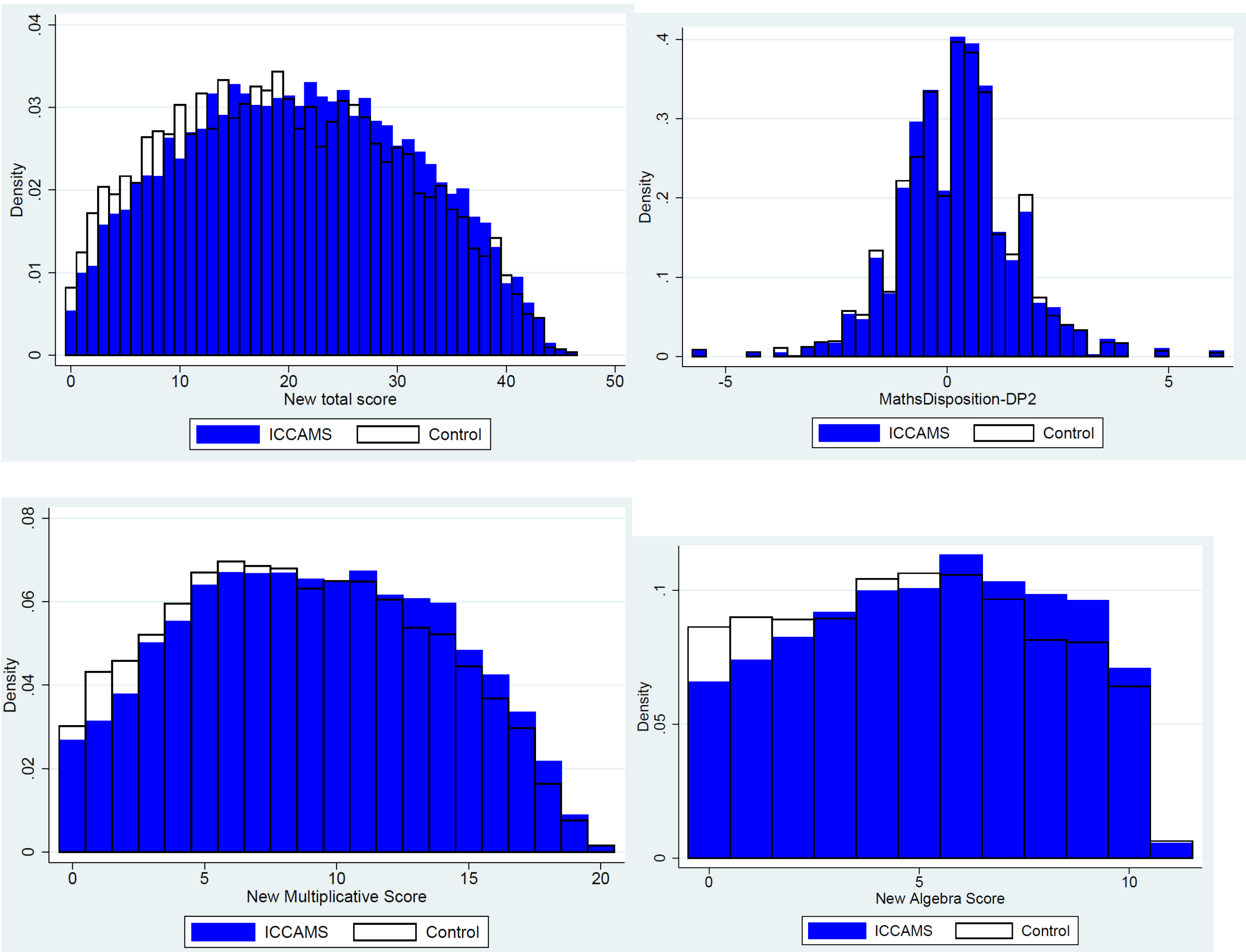


Figure 9: Histograms with the distributions of primary and secondary outcomes for intervention and control students (note: the maximum scores are 46 for the total, 20 for the multiplication, and 11 for the algebra subscale)

Analysis of outcome data followed intention-to-treat principles, for example, ignoring noncompliance, protocol deviations, and other events that take place after randomisation (Gupta, 2011). As randomisation took place at school level within each of the five regions—considering %FSMever and GCSEs as strata—this was taken into account via multilevel models to estimate a school-level and a student-level variance in order to allow for schools to differ regarding their average outcome. The unit of analysis was student-level outcomes and each model included the outcome of interest as dependent variable (that is, students' maths raw score and/or Rasch score) and the following covariates as independent/explanatory variables: an indicator of group membership ('condition', ICCAMS maths intervention vs control), blocking strata (region, with Region 1 as reference category, and FSM), and student's KS2 score—KS2_MATSCORE—as explained earlier (see equation 1):

$$Y_{ijPost}=\alpha_1+ \alpha_2.Cond_i+\alpha_3.Y_{ijPre}+ \alpha_4 Region2 + \alpha_5 Region3 + \alpha_6 Region4 + \alpha_7 Region5+\alpha_8.FSMever+\varepsilon_{ij} \quad (1)$$

Where:

Y_{Post} = standardised MALT mathematics scores (raw/Rasch scores)
Y_{Pre} = scaled KS2 mathematics attainment scores
$Cond_i$ = a dummy variable with the reference category indicating the control group
ε = error term for pupils clustered at school level
i = pupil i
j = school j

The coefficient α_2 associated with the condition dummy is the main result of the trial. All statistical models were performed in Stata 14 with the command `mixed`. See Appendix 15a for relevant code.

The output of these models includes the variances at school and individual student level, which were used to calculate the effect sizes as described later. The intra-cluster correlations (ICC) were calculated using Stata's post-estimation command (`estat icc`) for each model (example outputs are also shown in Appendix 15a for the primary outcome).

Secondary analysis

In order to respond to RQ2 and RQ4 we have employed the same modelling approach as with the primary outcome. For the secondary attainment outcomes (RQ2), two multilevel models were derived based on equation 1, but with different dependent variables as follows:

- secondary outcome model 1: Y_{Post}—attainment subscale of multiplicative reasoning (raw/Rasch scores); and
- secondary outcome model 2: Y_{Post}—attainment subscale of algebra (raw/Rasch scores).

Analysis was performed again based on both the raw and the Rasch scores, so for each outcome we have two set of results: priority is given to the raw score (for comparable results with other evaluation studies), which are presented in the main document and results with the Rasch scores are presented in Appendices.

For the secondary attitudinal outcome (maths dispositions) two models were derived based on equations 1 and 2 with dependent variable Y_{Post}—maths disposition at DP2. Two further models were derived to account for students' mathematics disposition at baseline (collected via surveys in September and October 2016) as denoted in equation 2:

$$Y_{ijPost}=\alpha_1+ \alpha_2.Cond_i+\alpha_3.Y_{ijPre}+ \alpha_4 Region2 + \alpha_5 Region3 + \alpha_6 Region4 + \alpha_7 Region5+\alpha_8.FSMever+ +\alpha_9.\text{MathsDispositionDP1}+\varepsilon_{ij} \quad (2)$$

Analysis in the presence of non-compliance

In this section we consider how school fidelity was measured and the analytical approach to dealing with non-compliance.

Capturing fidelity

School fidelity data was captured considering the three key aspects of the ICCAMS intervention, as agreed with the delivery and developer teams during the second year of the intervention: (1) attendance at PD sessions by the lead teachers, (2) the delivery of cascade training within schools during the two years of the intervention, and (3) the number of ICCAMS lessons reported as taught in the second year (Year 8). These elements were integrated into the teacher survey, which was administered at the end of the second year (June 2018—see Appendix 7).[25] Table 12 shows the specific items and their scoring as agreed with the developer and delivery teams with some further changes or clarifications that were added at the analysis stage.

Table 12: Fidelity items at school level along with some necessary changes that had to be implemented at analysis stage

	Low	Medium	High	Changes during analysis
Score	1	2	3	
PD Attendance	1–9 sessions	10–15 sessions	16–18 sessions	Between the two lead teachers
Lessons taught (at Year 8)	<15	15–17.4	17.5–20	Average as explained below
Cascade	No, missed more than 2	No, missed up to 2	Yes	

The major change in the above scoring related to capturing the number of lessons taught. The initial intention with the agreed plan, as presented in Table 12, was to measure this aspect at school level, but the practicalities of that were not detailed. Due to the variation we noted from teacher surveys, we opted for using all available information based on teacher survey responses. In particular, teachers were asked to report the class name and how many sessions they taught in the second year of the study (that is, to Year 8). In order to calculate the average lessons per school—as fidelity was to be a school-level measure—we went through a thorough cleaning and matching process for this dataset, including the following steps:

- removing duplicates in case a class was reported by two teachers;
- adding up the lessons taught by each teacher in cases where they were less than 20; and
- reducing to 20 anything reported that was above 20 (there were cases reporting 22 and 24, but the expectation was to teach 20 lessons; the extra lessons were designed for 'extra' help that some teachers might consider necessary for some classes).

That resulted in a unique class-level dataset for reported lessons taught that was used to average per school. The same information was further matched back to the student-level dataset for other analyses.

The three resulting items used to construct the fidelity measure, also correcting for minimum reported cascade in case of disagreement between the two lead teachers, were analysed with the Rasch Rating Scale model (Linacre, 2000).[26] More details are provided in Appendix 13 and some descriptive statistics are presented under the IPE results.

The fidelity scale was constructed at school level and validated using the Rasch model and our usual procedure based on data from 53 schools. Only eight schools met all the highest fidelity scores indicating expectations of number of lessons taught, lead PD attendance, and cascade implementation, while more than a third of the sample of schools were deemed low on all counts. Taking the items used by the team to evaluate fidelity at face value, we can see in Figure 10 that the schools fall into three main clusters separated by two boundary scores (at 1.1 logits and -0.75 logits) with two schools below the minimum level of -3 logits and four close to the edge of the highest boundary. In total, from these responses and as noted in Figure 10, 15 schools fall under each of the high and medium fidelity categories and 23 within the low fidelity category. The cut off scores then would correspond in raw scores as shown in Figure 11, which is consistent with the categories assigned by the total raw scores. The case study schools (for example, labelled 'S-84' for school 84) are also shown on this scale for cross reference with the IPE results.

[25] It should be noted that it was initially planned to collect information on lessons taught through emails to lead teachers by the delivery team and followed up during testing at schools by the evaluation team but this was also replaced by the information on teacher surveys. There was also a first year teacher survey run by the delivery team, however, the information from that was deemed not fit for purpose to capture fidelity.

[26] The natural decision for a model with items with different response options would be the Partial Credit model, However, for this case the Rating Scale model was the best option due to the small number of responses (52 schools), having only three items, and the fact that some response categories had fewer than ten observations.

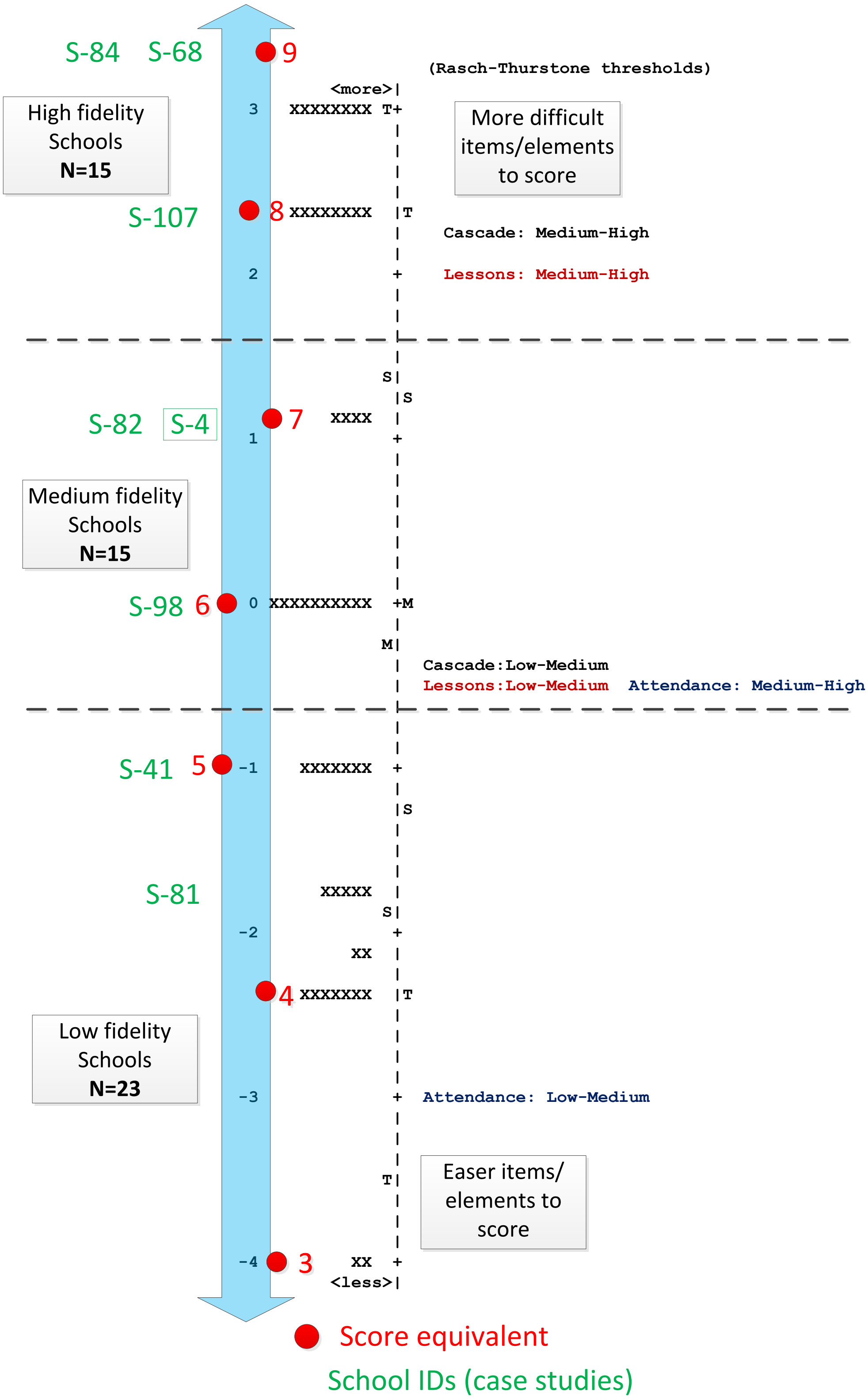


Figure 10: School-item map (Rasch-Thurstone Thresholds) for the fidelity scores

Modelling with the score of fidelity and considering non-compliance

We performed two types of models in order to check for the effect of fidelity as well as non-compliance.

The resulting measure was used to examine the relationship between fidelity and attainment outcomes in quantitative models (that is, in response to RQ5: Is there an interaction between fidelity and attainment change for the treatment schools?) but also as triangulating information in the IPE.

For the former, the following model specification was used (building from equation 1) considering 'fidelity' as a continuous variable (measured with the Rasch model as described earlier) and essentially was based on the sample from the intervention only:

$$Y_{ijPost}=\alpha_1+\alpha_2.\text{Fidelity}+\alpha_3.Y_{ijPre}+\alpha_4.\text{Fidelity}*Y_{ijPre}+\alpha_{5\text{-}8}Region2\text{-}5+\alpha_9.FSMever+\varepsilon_{ij} \quad (3)$$

In the protocol, it was mentioned that analysis will be performed with a class-level indicator of fidelity. However, as detailed in the previous section, that was not possible given the elements of fidelity that were agreed were referring to the school. However, class variation was captured via three-level models (both considering class or teacher). These are presented under additional analyses later (see Table 13).

Furthermore, in accordance with standard analytical procedures in EEF evaluations and our protocol—and in order to avoid underestimating the resulting effects—we employed an instrumental variable approach as it was considered more advantageous to the Complier Average Causal Effect (CACE) analysis (Gerber and Green, 2012). The instrumental variable approach was implemented via Two-Stage Least Square models (2SLS) (Angrist, 2006; Angrist and Imbens, 1995; Angrist, Imbens and Rubin, 1996) with group allocation as the instrumental variable. The compliance indicator was specified with three different variants considering the complexity around (non-)compliance. In fact, non-compliance could manifest in this study at various levels (student, class, teacher, or school) even though it was not possible for all to be observed: for example, we could not associate a student level-indicator as we did not have information from all classes about dosage (that is, how many ICCAMS lessons were taught) or even a full match of students to classes or teachers.

The 2SLS models were run with Stata's `ivregress 2sls` command. The first stage of an IV approach to compliance is prediction of the compliance indicator using the treatment allocation and then using the predicted IV plus all other predictors at the second stage to model the outcome. As multilevel specification is not available, we accounted for the clustering through robust standard errors (see Appendix 15b for model specifications and results). As a final compliance check, a model considering the interaction of the compliance indicator with the condition was also performed. It should be noted that for this analysis (as well as some of the explanatory analysis reported later), discussion is based on the coefficients in the models compared to ITT models.

Checking for imbalance at baseline

This was checked in relation to KS2 data (from the NPD) with difference being presented as an effect size. Comparisons in background characteristics (for example, gender, FSMever) between the control and intervention schools, and between regions, are also reported. Similar comparisons were performed with students' attitudinal baseline measures.

Sensitivity analysis

Outcomes from both analyses (with Rasch and raw scores) were compared to understand if the raw scores have violated assumptions of linearity and then to confirm whether there were any important differences in the model and the coefficients for the estimation of effect sizes.

Missing data analysis

Our approach to dealing with missing data was guided by the extent of, and the patterns of, missingness (that is, whether data was missing completely at random, at random, or not at random). To determine these we first report complete cases and establish the mechanism of missingness via multilevel logistic regression models where the probability of missingness was modelled (on the basis of responses and complete student lists shared by the schools) with additional predictors (including school and class level and other available information). These were performed with the `xtmelogit` command in Stata with the outcome of interest being whether the student missed the test (variable 'MissingTest').

Since missing (student) cases in intervention and control groups (also considering school attrition) were greater than the maximum 5% threshold considered safe for bias as per EEF's guidelines, analysis proceeded depending on the mechanism of missingness as defined in our SAP.

Sensitivity analysis was also executed by comparing a complete case analysis to multiple imputations for the MLM models performed using Stata. Thus we were able to include partially observed cases—cases that have not got a value for each of the variables in the model. In particular, we created ten imputed datasets using multiple imputation with chain equations with the command `mi impute chain`, which iteratively imputes missing values according to the pre-specified format of the variable (that is, logistic model for binary variables and linear model for continuous). We have accounted for the hierarchical structure of the data by including the school variable as a covariate in the imputed models (see Appendix 15d, also note: other attempts to account for clustering by separate imputation per school were not converging).

The resulting ten imputed datasets were then combined into one set of results with the command `mi estimate: mixed` (with the same models as presented with equations 1 and 2 earlier under ITT analysis, see Appendix 16).

Subgroup analyses

Subgroup analyses were performed to answer RQ3. The effect of the intervention on attainment (primary and secondary outcome measures) was analysed by repeating the primary analysis for the subgroup of students who are eligible for FSM. In other words, only pupils eligible for FSM were selected to form a new subset. The following models were thus specified for each outcome, as explained for the primary and secondary outcomes, earlier:

$$Y_{ijPost}=\alpha_1+ \alpha_2.Cond_i+\alpha_3.Y_{ijPre}+ \alpha_{4\text{-}7}.Region_{2\text{-}5}+\varepsilon_{ij} \qquad (4)$$

Additional analyses and robustness check

In order to explain and understand more deeply the effects of the ICCAMS intervention we also ran additional models with further covariates at student and school levels, which have been shown in previous literature to mediate or moderate such primary and secondary outcomes (these are detailed as explanatory analyses in the SAP). The general form of these models to account for these measures are based on equation 1 and extended to equation 5:

$$Y_{ijPost}=\alpha_1+ \alpha_2.Cond_i+\alpha_3.Y_{ijPre}+ \alpha_{4\text{-}7}.Region_{2\text{-}5}+\alpha_8.FSMever+bX_{ij}+\varepsilon_{ij} \qquad (5)$$

This is defined as in equation 1 but with the addition of *X* as the vector of other related covariates (and their associated vector of *b* coefficients).

We thus replicated the models as in equation 1 but now considering gender, age, and the measures of students' dispositions and perceptions of transmissionist teaching and its interaction with the intervention indicator. We further checked for the class/teacher level effect—also with covariates related to the teacher such as their perception of transmissionist teaching, perception of teaching for FA, and confidence teaching through ICCAMS, (the latter, as denoted with the set of models under M5 and M6 in Table 13, were only measured with the intervention teachers).

Additional models were run with a new categorical variable to denote the type of teacher at allocation (lead or cascade). The models explored are summarised in Table 13.

In addition to the pre-specified explanatory models in Table 13 we also ran models to explore teachers' level of self-reported transmissionism and changes in learners' perception of transmissionist classroom practices as outcome variables. In particular, we model students' perceptions of transmissionist teaching practice as an outcome, with explanatory variables those in the main ITT models, and important variables found from Models M1-M6 stated in Table 13 along with other available and relevant variables, for example, perceived difficulty of the lessons and the interaction of intervention condition and type of teacher. A model of average mathematics dispositions at DP2 (of all the students assigned to each teacher with a provided response) considering (1) average mathematics disposition at DP1, (2) teacher type (lead/cascade), (3) years of teaching experience, and (4) teachers' perceptions of FA practice and transmissionist teaching at DP2 was also run as triangulation of the IPE findings. This was an amendment to the original protocol motivated by results of the IPE that suggested that such changes had likely occurred in the intervention schools' classrooms.

Table 13: Overview of models run for explanatory analysis

Set of Models	Outcome	Extra Covariates Base models: Condition + KS2Math +FSMever+Region +	Model Structure
M1 M1a	Attainment Algebra Multiplicative Math disposition	Age +Gender +Teacher Type	2-level (students in school)
M2 M2a	Attainment Algebra Multiplicative	Age+ Gender+Math disposition 1 +Teacher Type	2-level (students in school)
M3 M3a M3b	Attainment Algebra Multiplicative Math disposition	+ Age+ Gender+TransTeaching 2 +Teacher Type (also control for teaching at baseline by including TransTeaching1)	2-level (students in school)
M4 M4a M4b	Attainment Algebra Multiplicative Math disposition	+ Age+ Gender+TransTeaching2+ Teachers FA+TeachersTrans +Teacher Type (also control for teaching at baseline)	3-level (students, classes/teachers, schools)
M5 M5a	Attainment Algebra Multiplicative Math disposition	Age+ Gender+TransTeaching 2+TeacherTrans+TeachersFA +Confidence Teach ICCAMS +Teacher Type	3-level: ICCAMS intervention data only (students, classes/teachers, schools)
M6	Attainment Algebra Multiplicative Math disposition	+Fidelity interactions	3-level: ICCAMS intervention data only (students, classes/teachers, schools)

Estimation of effect sizes

Effect sizes are reported as Hedges' g and were calculated based on the adjusted mean difference between the intervention and control group (that is, the coefficient of the 'condition' variable in regression models controlling for prior and other randomisation stratifiers with the models specified with equation 1) and the variance components produced by Stata 14.

Effect sizes are accompanied by 95% confidence intervals as per EEF recommendations using two-sided tests at the 5% level unless otherwise stated. In particular, the 95% confidence interval for the ES that took into account the clustering of pupils in schools was calculated by dividing the upper and lower confidence interval bounds by the variance components as noted above. Estimates of effect with 95% confidence intervals (CIs) and p-values are provided as appropriate.

Implementation and process evaluation

Research methods

The proposal and protocol set up the terms for further data collection through a process evaluation with an aim to isolate the causal explanations of the intervention impact under usual conditions. Our implementation and process evaluation (IPE) was thus focused on both collecting evidence about how the intervention was conducted at its various levels (at PD, schools, and classrooms) as well as how this might affect and explain the hypothesised primary and secondary outcomes and the relationships in the statistical models. In particular, we collected and analysed evidence around fidelity and integrity (RQ6), dosage and exposure (RQ6b), participant responsiveness (RQ8), possible adaptations (RQ7), and some evidence for the 'business as usual' practices in control schools (RQ9). The process evaluation also provides opportunities to explore teachers' attitudes and practices (RQ10) as well as the effect of cascading (by comparing cascade vs lead teachers). The main highlights of our approach to IPE are overviewed below whilst Table 14 summarises the collected evidence and its association with the research questions.

Teacher surveys

The mathematics teachers in the randomised sample of schools were a participating group of interest. According to the logic diagram, a relevant subgrouping for the intervention group regards the teacher's level of participation in PD: direct (for 'lead teachers') or via school cascading (for 'cascade teachers'). Teacher knowledge, beliefs, and perceptions of practices were collected through teacher questionnaires at the start (September/October 2016) and end of the project (May/June 2018) in all control and intervention schools. This was deemed crucial for the evaluation of this intervention as it would provide an indication or indicative measures of what was happening during the mathematics teaching in order to objectively monitor practice and measure variations in the degree of cascade effect, which was expected to vary from school to school. We hypothesised that involvement with ICCAMS PD will have an effect on generic aspects of a teacher's practice, which was captured via a measure of teachers' perception of 'transmissionist' teaching. In addition, we also hypothesised that there will be more direct effects on teachers' practice of formative assessment so we extended previous instruments developed and validated by the evaluation team (Pampaka and Williams, 2016; Pampaka et al., 2012b) with more items focusing on this aspect, drawing on details of how ICCAMS PD was delivered and other related work (Herman, Osmundson and Silver, 2010; Wiliam, 2007a). We added such items to our previously validated transmissionist scale hypothesising that there could be a second dimension for practice more related to FA (as detailed in the earlier section on Measurement, see also Table 11).

Student survey

The student survey questionnaires (apart from the mathematics disposition scale) also included an instrument measuring students' perceptions of the teaching they receive (see Appendix 7 for whole questionnaire). This was used as a moderator/mediator in some models for primary and secondary outcomes (in a similar manner to those of the teacher survey measures). We found in previous work that this measure can be a significant explanatory variable in models of students' dispositions towards mathematics, and sometimes more significant than teachers' more subjective self-reported views on practice. In addition, capturing this information from students was designed to mitigate the risk of (teacher) missing data as usually teachers are less likely to complete such surveys.

Observation of professional development and PD lead training events

We observed five PD events attended by lead teachers in the first year and five in the second year, that is, once in each of the five regions and with a combination of joint observations and researchers attending various PDs for comparability (see Table 14). We also attended three 'training the trainers' days for the PD leads in the first year: two investigators attended the September 2016 (first) session via Skype and a researcher attended the November 2016 and March 2017 events during which the lead evaluator also joined via skype.

Table 14. Observations of PD sessions by the evaluation team

	Region 1-EA	Region 2-EM	Region 3-L	Region 4 SW	Region 5- Y
PD 1		R1 and R2			R1 and R2, PI*
PD 2			R1		R1
PD 4					R1
PD 8	R3	R3	R3	R3	R4
R1 to R4: codes for four researchers involved in this task over the two years. PI*: principal investigator of evaluation, attended via video link.					

Field studies in schools

Field studies involved interviews with stakeholders (PD leads, teachers, students) and lesson observations with lead and cascade teachers. (Teachers were also observed at some of the PD training events as noted above.)

At the protocol stage, we proposed to choose our sample of schools by selecting at least two after advice about diversity from each PD lead and considering the fidelity measures. However, for logistical reasons the fidelity measures were constructed after data collection. Challenges with getting responses and invitations from schools—especially regarding student interviews—resulted in fewer visits and interviews (nine individual interviews) than initially planned, however, we visited eight schools representing every region, interviewed 21 teachers, and observed 19 lessons (Table 15). In Table 15 we also report the retrospective classification based on the school fidelity scores, which shows that we achieved a range of schools with diverse fidelity and in particular we had two maximum fidelity schools included. We had also considered average teacher scores in FA practice and transmissionist teaching upon selection of interesting cases, however, as noted above, the final list was guided by school's availability and willingness to be visited at the time we requested such access. We gathered a richer than expected IPE dataset due to connectivity of these schools with rich survey data (see Table 16). These data analyses sought explanations including the ways institutional and cultural norms, pedagogy, and ICCAMS materials and training mediate teaching and learning practices and outcomes. Despite the structured agenda for data to be collected from these sites, we kept these studies as open ended as possible including being open to the unexpected in schools that seemed different, or otherwise interesting. The purpose of these studies was to test the ICCAMS 'hypothesis' that the ICCAMS intervention can cause improvements to teaching and learning. This necessitates explanations for phenomena, whether evidenced or just anticipated (for example, differences between lead and cascade teachers) or unanticipated initially.

Table 15: Summary of data collected from case study schools

School ID	Region	Teacher interviews	Student interviews	Observations Year 1	Observations Year 2	School Fidelity Scores*
4	Region 2	2L + 3C			1L + 1C	1.08 (medium)
41	Region 3	1L + 2C		1L	1L + 1C	-0.99 (low)
68	Region 4	[1L + 1C]***			1L + 1C	3.85 (highest)
81	Region 1	1L + 1C			1L + 1C**	-2.35 (low)
82	Region 5	1L + 1C	9 students	1L + 1C	1L + 1C	1.08 (medium)
84	Region 2	1L + 2C			1L + 1C**	3.85 (highest)
98	Region 2	1L		1L	1L	0.04 (medium)
107	Region 5	1L + 4C			1L + 1C	2.33 (high)

Notes: L = lead teacher; C = cascade teacher.
* The final column of the table, the result of analysis, is included so the reader can see the range of fidelity scores.
** Lessons observed with cascade teachers were non-ICCAMS.
*** Interviews were not recorded; only notes were taken.

Control school surveys

As per the protocol, in order to monitor the control conditions we also collected further information at the end of academic year 2017/2018 from control schools and control school teachers through surveys about their experience and any actual PD going on in their schools. The teacher surveys (as explained earlier) also provided wide information on teachers in these schools.

Analytical approach

IPE was based on analysis and synthesis of the data described in Table 16.

Table 16: IPE methods overview

Research methods	Data collection methods	Participants/data sources	Data analysis methods	Research questions addressed	Implementation/logic model relevance
Researchers observations at PD sessions	Observation field notes	PD trainers and teachers (n = 1)		RQ6	Fidelity, programme differentiation, adaptations
Researchers observation of PD training Sessions	Observation field notes and material presented	PD trainers (n = 8)		RQ6	Fidelity
Researcher lesson observations (intervention)	Lesson Observations along with PD leads	Teachers and pupils (n = 4)		RQ6	Fidelity
Interviews with PD leads	Interview Transcripts	PD trainers (n = 4)	Thematic and reflective cross case analyses	RQ6	Fidelity, quality, adaptations
School case studies	Classroom observations	Teachers and pupils/ researcher fieldnotes (n = 19 lessons)		RQ6 to RQ10	Quality, fidelity, programme differentiation
	Teacher interviews	Teachers (n = 21)		RQ8 RQ7 RQ6	Responsiveness, fidelity, quality, reach
	Student short interviews	Pupils (n = 9)		RQ8	Responsiveness
Teacher survey (questions for All)	Teacher questionnaires	Teachers (DP1 n = 596) (DP2 n = 424)	Quantitative analysis (descriptive and modelling)	RQ10 RQ7	Programme differentiation, monitoring of control/comparison groups
Student survey	Student questionnaires (perceptions of teaching practices)	Pupils (DP1, n = 17644) (DP2, n = 18115)	Quantitative analysis	RQ10 RQ7	Programme differentiation
Teacher survey (questions for intervention only)	Fidelity related questions	Teachers (DP1, n = 296) (DP2, n = 204)	Surveys; quantitative analysis	RQ6	Fidelity, dosage
Email survey (control schools)	Open questions to control school key contracts	School-level responses (n = 24)	Thematic analysis	RQ9	Monitoring control/business as usual conditions
Lesson observations and schedules by PDs	Observation records by PD leads and delivery team	Lesson observations (Year 1, n = 113) (Year 2, n = 77)	Secondary data	RQ6	Fidelity (school fidelity measurement)
Teacher surveys	Teacher Surveys with intervention teachers (collected by delivery team during Year 1)	Teachers (LT, n = 90) (CT, n = 137)	Secondary data	RQ6	Fidelity, dosage, reach

Notes:
DP1 refers to the start of the evaluation (start of Year 7) and DP2 refers to the end (end of Year 8).
Reflective case studies involve testing the data at school level and then at multiple school level against the questions about causations in relation to the impacts being measured. The reflection is also supported by relevant literature as well as our experience of the limitations of this study.

There were several stages in this analysis involving both thematic analysis of qualitative evidence across the teacher interviews and a reflective case study analysis at the whole school level and whole sample level as well as quantitative analysis at various levels based on school-level data, teacher-level data, and student-level data. This included both descriptive presentation of findings with summary tables and figures as well as further models for explanatory purposes. Details of analysis are also provided along with the corresponding results under the IPE results reported below.

More is provided here about (1) the thematic analysis of the teachers' interview data and (2) the whole-school, whole-sample studies. The thematic analysis of predominantly interview data followed an empirically grounded method where readings of the data were used to develop categories to exhaust and subsume the data through a finite number of themes that were arising across the successive interviews and observations. Comparisons are made taking account of variables such as 'lead' versus 'cascade' experiences and other factors raised by the teachers.

Bearing in mind this analysis, the data was later re-read at school level in the light also of the literature review and its conclusions about PD in these contexts. The possible explanations for the intervention being more or less successful were explored though readings of all the data available on each of the schools. Of course, most of the findings of the thematic analysis were here validated, but it was possible to discern some features of what appeared to be more successful school-level practices and their cultural conditions (all these are reported later). We were also able to use some of the survey data in this synthesis and the result is a credible consideration of explanations for the factors that matter in the implementation of the intervention and why impacts may have been significant or attenuated.

In particular, the logic model in the protocol (and as shown in Figure 4) guided our analysis at school level: we were able to compare high fidelity schools with low fidelity schools and use the data to 'explain' similarities or differences by pointing to causes of breakdown of the model that distinguish or that prove common to both types of school. 'Explanations' in such work should arise as causes that are credible rather than 'frequent' as such, but the frequency is still worth noting. When we looked at the data, then, we did so across the methods using the school as a unit as well as considering the connectedness of data at teacher and class level (that is, from observations, interviews, as well as surveys).

Costs

Cost information was based on the costs of training provided by the developer and delivery teams on the assumption that the intervention is run as intended (that is, two teachers per school attend and so forth). Calculations were based on the December 2019 'Cost Evaluation Guidance for EEF Evaluations'.

Data was collected from the developer and delivery teams as well as directly from schools to uncover the expected and any unexpected costs of this intervention. In particular, the delivery and developer teams provided the following information:

- PD lead fees for PD delivery—including preparation for delivery and delivery;
- PD lead fees for school lesson observations;
- PD lead travel to PD and observations costs;
- venue and catering costs—lunch and refreshments provided; and
- materials—including initial handbooks and revisit materials.

As part of the evaluation, we collected data regarding the following costs for schools implementing their responsibilities in the intervention, as well as related outside school costs:

- direct, marginal costs—costs directly attributable to the school's participation in the intervention; including in relation to providing the training, school visits (if necessary), implementing the intervention, and providing resources for successful implementation; in particular, we aimed to account for the costs related to salary (for teachers' time to take part in training), printing, and fees for services;
- 'hub'-related costs of administration and implementation monitoring; and
- prerequisites, especially in relation to the delivery of the intervention (and regarding the equipment needed or available at schools).

A further school survey run at the end of Year 2 via emailing asked intervention schools about the following:

- Lead teacher: how much additional time per week would a teacher spend preparing for ICCAMS lessons?
- Cascade teacher: how much additional time per week would a teacher spend preparing for ICCAMS lessons?
- What was the financial cost for teacher cover (with two teachers attending a professional development day)?
- What were the travel and subsistence costs when attending a PD day?
- Any other financial costs of ICCAMS participation (for example, printing) for the school?
- How many of your students were involved in the ICCAMS intervention 2016–2018?

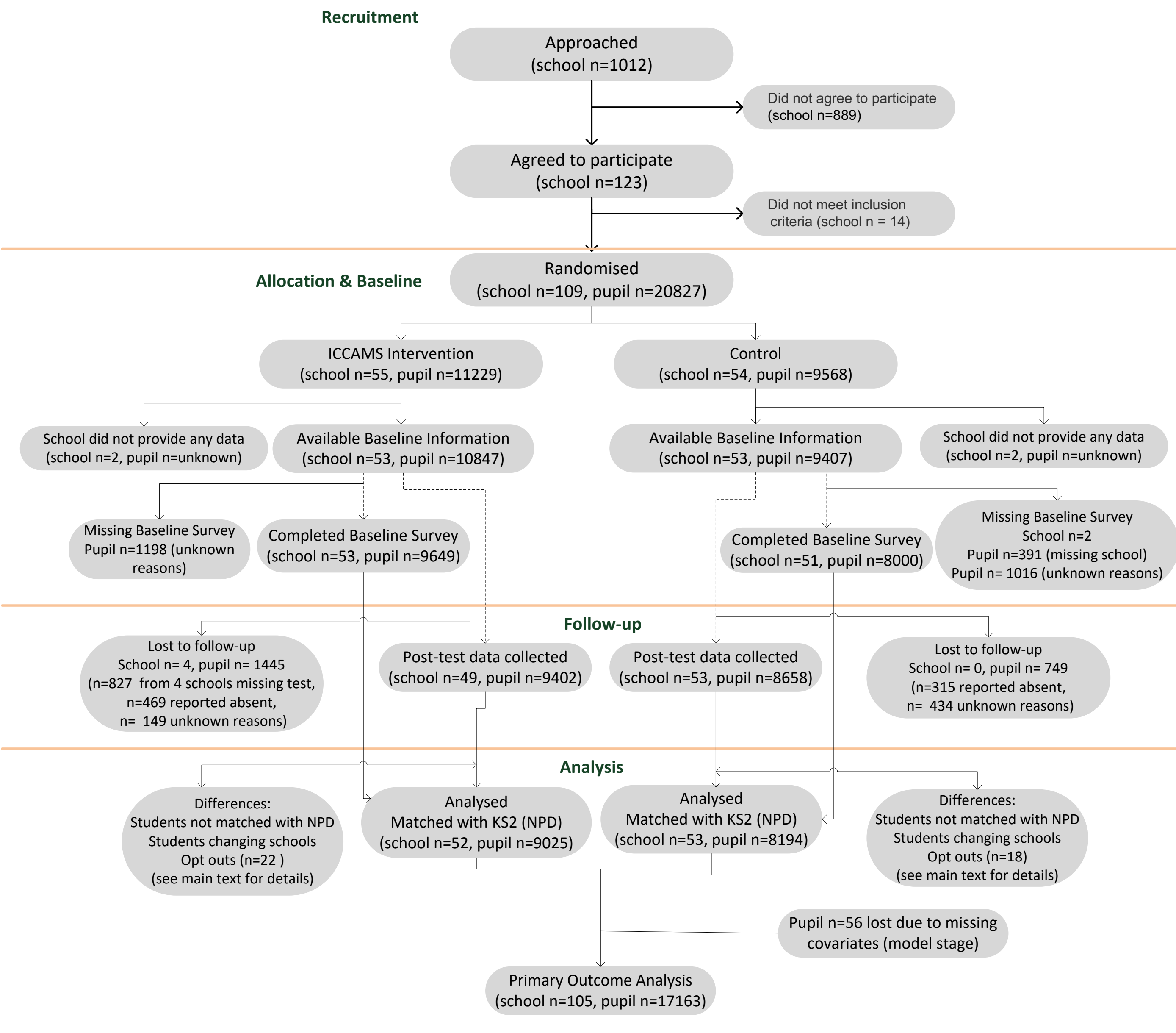
Timeline

Table 17: Timeline

Dates	Activity	Staff responsible / leading
1 Aug 2015	Project starts	
Aug 2015–Jan 2016	Intervention development work	
Oct 2015–Aug 2016	Development of new instruments and institutional ethic clearance	All
Jan 2016–Jul 2016	Piloting of intervention, training and materials	Developer/delivery teams
June–Aug 16	Validating new instruments and subscales	Evaluation team
Jan–Jun 2016	Recruitment of schools to trial	Developer/delivery teams
Oct 2015–May 2016	Recruitment of PD leads	Developer/delivery teams
May 2016	Training of PD leads	Developer/delivery teams
July 2016	Randomisation of schools	Evaluation team
Sep 2016	Pre-survey First year of intervention begins	Evaluation team
Sep 2016–Jul 2018	Process evaluation	Evaluation team
Sep 2017	Second year of intervention begins	Evaluation team
Jun–Jul 2018	Final outcome assessment	Evaluation team
Sep 2018–Aug 2019	Re-visiting data sharing agreements Class data shared August 2019	All
Sep 2018–Aug 2019	Data entry, cleaning, matching	Evaluation team
Sept 2020–Feb 2020	Access to ONS /NPD Draft report	Evaluation team
	Final report	Evaluation team
	Project end date	Evaluation team

Impact evaluation results

Participant flow including losses and exclusions



Note: there is not an exact baseline student sample frame at randomisation, as is discussed further in the Pupil Characteristics section below. For these purposes, we report our best estimate of the retrospectively calculated student sample in order to show their flow through the study. Where lines are dashed the flow is not based on the same matched information.

Figure 11: Participant flow diagram

The minimum detectable effect size at different stages of the evaluation is presented in Table 18 recalculated with the calculator tool as described earlier. The average cluster size was larger than originally assumed (as discussed below) and the ICC for school level is lower than expected (0.07 instead of 0.12). This means that the sample size and thus the power of statistical tests would be greater than expected.

Table 18: Minimum detectable effect size at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.15	n/a	0.19	n/a	0.146	0.103
Pre-test/post-test correlations	Level 1 (pupil)	0.65				0.79 (raw) 0.8 (Rasch)	0.74 (raw) 0.76 (Rasch)
	Level 3 (school)					0.74 (raw) 0.78 (Rasch)	0.72 (raw) 0.76 (Rasch)
Intracluster correlations (ICCs)	Level 3 (school)	0.12		0.12		0.07	0.04
Alpha		0.05	0.05	0.05	0.05	0.05	0.05
Power		0.8	0.8	0.8	0.8	0.8	0.8
One-sided or two-sided?		2	2	2	2	2	2
Average cluster size		75-100		190.1		171.9	47.4
Number of schools	Intervention	50		55		52	52
	Control	50		54		53	53
	Total:	100		109		105	102
Number of pupils*	Intervention	5000		11229		9396	2513
	Control	5000		9568		8656	2468
	Total:	10000		20827		18052	4981

* Numbers reported here are the smaller estimated for the randomisation and the available cases with the primary outcome for the analytical sample. Even though only 102 schools completed tests at the end of the study, there were three additional schools with observed student data due to student movements. For further missingness patterns see other parts of the report.

Pupil and school characteristics

Sample sizes are presented in Figure 11 and Tables 19 to 22. The ICCAMS delivery team approached 1,012 schools to participate in this trial but the majority did not express interest in participating (as indicated mainly through not responding positively to the invitations). From those 123 that agreed, 14 were not eligible and therefore the resulting 109 schools that fulfilled the criteria and signed the school agreement were randomised into 55 trial and 54 control schools.

The precise number of students involved is not knowable as randomisation took place before the start of Year 7 and was based on a school list (including the recruited 109 schools). This list of schools did not have any information on expected Year 7 intake; information received at later stages of the evaluation was also incomplete and could not provide an accurate estimate of the sample at randomisation. In particular, the pupil data provided by schools was not available in full for all schools (four schools did not provide any pupil data and others were missing other information such as UPNs) and in some cases there were large differences with the lists for pupils starting in Year 8. Further attempts to get an accurate estimate from other school-level information (using NPD data as well as open data) did not provide specific breakdowns by year group (there are variables available on school capacity or full school cohort sizes). This may also explain the difference between the Year 7 size difference we used at the protocol for the power calculations and the observed averages at the other stages (Table 18).[27] When we received matched NPD information from the school census, information on pupil numbers was only available by pupil age, not by year group.

Therefore, we constructed a 'best estimate at baseline' using the three-step process as described below:

- step 1—the primary source was the Year 7 cohort size as stated in the school lists (provided by schools);
- step 2—information collected at baseline from student surveys was added: in one case when information was not provided by the school, the survey sample was added to the total, and in two cases the school size was replaced with the survey sample as it was larger than the list; this resulted in information available from 106 schools involving 20,254 students (average 191 per school); and
- step 3—the average Year 7 cohort size based on the available information from steps 1 and 2 was used as an approximation for the three missing schools resulting in a total of 20,287 students as the best approximation for baseline.

It should be noted that this baseline estimate is only based on information available from separate school lists, before any attempt to match with final data. The flow from the baseline information to follow-up and analytical sample is thus given with dashed lines to avoid unjustified assumptions (that is, that the various stages are based on exactly the same cases). We also added arrows from baseline survey sample to the analytical sample to indicate the further reduction of sample size for some models which draw on this information (for example, mathematics disposition at DP1). We believe that this is a good estimate of the baseline sample size, however, given that this information was not always possible to match with the analytical sample (for reasons we explain below) we use our 'best estimate for baseline sample' and urge for caution when one looks at the missing data information in the tables with the results that follow. To make more sense of further issues with this as a baseline estimated sample size, the reader should also consider the following information on the resulting (matched) sample.

'Retrospective baseline' refers to all cases that we could match in our database[28] and in particular includes information from DP1 (the student survey) and DP2 (test and survey), some of which was unmatched with NPD data and pupil lists provided by schools (at different stages of the matching process) and after removing 40 cases with opt-out or known duplicates (18 from control schools and 22 from intervention); our baseline total was 21,661 from the 106 schools that completed data collection activities. From these cases, 18,052 had valid test data and were thus the basis of the analytical sample which is presented in Table 18 (before losing further cases due to missingness in other covariates, such as NPD KS2 and FSM).

Before considering comparisons at baseline, it is useful to consider various further challenges in understanding the resulting samples, as well as the attrition patterns.

We start from school-level information noting that three schools were not responsive after randomisation; they are described as 'dropouts' in further sections. These schools did not provide any information at any stage (one control, ID 11, and two intervention with IDs 108 and 109), which means that on our baseline student sample we only have 'some' information from 106 schools. As noted above, we have provided an approximate sample size for these schools based on the average from the rest of the schools with some information. When it comes to school-level information, however,

[28] A further challenge was related to the delay in getting access to some of this information on the duration of the study, which could have enabled us to check back with schools—in fact the delivery team was only able to share pupil lists in August 2018. The preliminary power calculations (performed by the evaluation and the delivery team at set-up and first version of the evaluation protocol) was under the assumption of 75 -150 pupils per cluster but in reality the requirement for a minimum of three form entry meant cluster sizes were much larger.

especially for the key randomisation stratifiers (GCSE and FSM), we have information from these dropouts but for some other schools information was not available. At the testing stage (main outcome) four additional schools did not participate so the school sample at this stage is 102; however, some further complexities due to student-level movements and missingness are discussed next, as these make the analytical sample to relate to 105 schools. Various patterns regarding the stratifiers are shown in Table 19.

As shown, the changes in the proportion of FSM and GCSEs for both intervention and control schools are very small and should not be expected to bias any findings. What is perhaps less trivial are the potential implications of the missing information from other schools (for reasons beyond our control as simply they were not available within official tables).

Beyond these stratifiers, when looking at the school characteristics and their comparison with national averages, when available, in Table 21 we note that the control schools are more similar to the averages than intervention schools.

A final note on school attrition before we move to discussing pupil-level data. Further to the three school dropouts post randomisation that did not engage at all with the study, there were six more schools that withdrew from the intervention at some point—the four schools we noted above as missing test data and two other schools that despite dropping out either at the end of the first year (School 1) or mid Year 2 (School 16) participated in both baseline questionnaire completion and outcome testing. This information will be considered further when discussing compliance (and modelling in the presence of non-compliance).

Table 19: School-level demographics description

	FSM (school sample size and averages)			GCSE (school sample size and averages)		
	All schools (109)	Excluding dropouts (106)	Test completed (102)	All schools (109)	Excluding dropouts (106)	Test completed (102)
Control	54	53	53	48	47	47
Intervention	53	51	47	52	50	46
Total	107	104	100	100	97	93
Control	28.15	27.93	27.93	56.38%	56.47%	56.47%
Intervention	28.82	28.69	28.32	57.37%	57.64%	58.07%
Total	28.48	28.30	28.11	56.89%	57.07%	57.26%

When it comes to student data, the information is more complicated as there are various combinations of dropout and missingness. In the attrition Table (Table 20) we take the estimated baseline and contrast with the analytical sample, which includes those cases that were possible to match with KS2 results (17,219). However, we should note that some results are based on different analytical samples: for example, raw comparisons of means are based on the larger sample, which includes the available 18,052 who completed the test (after excluding opt-outs, and as presented in Table 18). In contrast, for some analytical models that consider responses to the DP1 student survey, the sample size drops further to 14,819.

It should also be noted that there were further missing data issues beyond the opt-outs: information not possible to match, schools not taking part in test or survey at DP1, students changing schools, and partially missing information on covariates. These issues are discussed under missing data analysis and compliance later on.

Table 20: Pupil-level attrition from the trial (primary outcome)

		Intervention	Control	Total
Number of pupils	Randomised (best estimated baseline)	11229	9568	20827
	Analysed (cases with tests matched with NPD)	9025	8194	17219
Pupil attrition (from randomisation to analysis)	Number	2204	1374	3578
	Percentage	19.6%	14.4%	17.2%

Comparison of key variables at school and pupil level at baseline are provided in Table 22 whereas Appendix 16 presents the comparative distributions of the two baseline measures (attainment and disposition at DP1) and the other secondary outcomes. As shown in Table 22, student-level characteristics indicate equivalence at baseline: apart from some small disparity in gender proportion (with proportionally more males in the intervention group than the control group), FSM proportions are very close as well as the baseline continuous measures of KS2Math scores and the measure of mathematics disposition. The larger difference is observed in regards to sponsored academies and urbanisation at school level when comparing the intervention with the control schools. It should also be noted that in contrast to the ideal design scenario, at school level, the proportion of FSM was slightly lower than the national average based on information at randomisation. However, when looking at the student level (analytical sample), this proportion is a bit higher than the average (30% for intervention and 31% for control schools). Some further checks have been performed in relation to the combination of variables crucial for the main models that will guide this analysis, such as the comparisons presented in Table 21. This indicates some regional differences in the distribution between control and intervention schools' FSM proportions with potential implications for the results about regional differences between Region 3 and the Region 4. But we judge these to be minor, perhaps offering some nuance to the later report of regional differences.

Table 21: Mean regional information on KS2Math and %EverFSM

		KS2Math		%EverFSM	
Region		Control	Intervention	Control	Intervention
Region 1	Mean (SD)	101.68 (7.11)	102.11 (7.11)	0.27 (0.44)	0.25 (0.43)
	N	1,714	1,978	1,797	2,042
Region 2	Mean (SD)	101.73 (6.81)	102.65 (6.85)	0.32 (0.47)	0.31 (0.46)
	N	2,413	2,483	2,583	2,580
Region 3	Mean (SD)	102.85 (6.75)	104.00 (6.78)	0.43 (0.5)	0.37 (0.48)
	N	1,802	1,731	1,870	1,835
Region 4	Mean (SD)	101.76 (6.54)	101.66 (6.45)	0.31 (0.46)	0.35 (0.48)
	N	1,399	1,901	1,452	1,974
Region 5	Mean (SD)	102.88 (6.97)	102.67 (6.8)	0.23 (0.42)	0.24 (0.43)
	N	1,987	2,790	2,063	2,859

Table 22: Baseline characteristics of groups as randomised

School level (categorical)	National averages	Intervention group		Control group		
		n/N (missing)	Percentage	n/N (missing)	Percentage	
Region: Region 1 Region 2 Region 3 Region 4 Region 5	NA	55/55 (0) 10 12 10 10 13	 18.2% 21.8% 18.2% 18.2% 23.6%	54/54 (0) 11 12 9 9 13	 20.4% 22.2% 16.7% 16.7% 24.1%	
School type: Academy converter Academy sponsor led Community school Foundation school Free schools/voluntary aided/voluntary controlled	 46.4% 20.9% 12.7% 6.2% 13.8%	55/55(0) 26 7 11 7 4	 47.3% 12.7% 20% 12.7% 7.3%	 22 12 9 4 7	54/54 (0) 40.7% 22.2% 16.7% 7.4% 13%	
Rural/urban location: Rural town and fringe Rural town+ fringe in a sparse setting Urban city and town Urban minor/major conurbation		55/55(0) 7 6 24 18	 12.7% 10.9% 43.6% 32.7%	54/54(0) 7 5 30 12	 13% 9.3% 55.6% 22.2%	
Religious denomination: No religious character CofE/Roman Catholic/other Christian	 81.4% 18.6%	55/55(0) 49 6	 89.1% 10.9%	54/54(0) 44 10	 81.5% 18.5%	
Ofsted ratings: Outstanding Good Requires improvement Inadequate No information (1 new school)	 22.3% 53.1% 16.8% 7.8%	52/55(3) 10/52 33/52 9/52 0 (1new+2)	 19.2% 63.5% 17.3%	50/54(4) 11/50 29/50 10/50 0 (4)	22% 58% 20%	
School level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	
Percentage eligible for FSM	29.1%	53/55(2)	28.82% (16.6)	54/54(0)	28.15% (16)	
Average of GCSE		52/55 (3)	57.4% (12.9)	48/54(6)	56.4% (13.5)	
Pupil-level (categorical)		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
Gender: Male Female		11572	6133 (53%) 5439 (47%)	10071	4943 (49%) 5128 (51%)	
Ever FSM No Yes	 71.9% 29.1%	11304	7921 (70%) 3383 (30%)	9780	6752 (69%) 3028 (31%)	
Pupil level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size
KS2 score (all NPD data)	102.87	10883 (346)	102.60 (6.84)	9315 (253)	102.18 (6.87)	0.045 (-0.05, 0.14) p=0.335
Mathematics Disposition DP1		9614 (1615)	0.55 (1.45)	7946 (1622)	0.56 (1.48)	0.012 (-0.07, 0.09) p=0.755

Outcomes and analysis

The evaluation assesses the effect of the ICCAMS intervention on mathematics attainment and attitudinal outcomes (primary and secondary outcomes), which are summarised with the histograms in Appendix 16.

The estimated difference between outcomes in the intervention and control group is based on an intention-to-treat analysis. This implies that all control schools are included, regardless of whether they participated in any relevant intervention during the trial, and all intervention schools are included regardless of whether they have completed the requirements of the trial. Further analysis for robustness and considering the effect of non-compliance are discussed below after the presentation and discussion of primary and secondary outcomes. We discuss each outcome below.

Primary analysis

Table 23 summarises the comparison of the scores on the primary outcome of the intervention and control schools and the calculated effect sizes based on the results of the multilevel regression model as specified with equation 1 (see Methods).

Even though there appears to be a notable difference in both the raw scores (as well Rasch scores: see Appendix 16) on the primary measure of mathematics attainment when comparing intervention with control group means, when controlling for previous attainment and other variables related to randomisation, the effect of the condition disappears. Students in ICCAMS intervention group make no more progress in mathematics as compared with the control schools, with a high chance that this effect was null. The estimated effect size confidence interval ranges from -0.07 to 0.15, which corresponds to between one month less to two months more progress of the intervention students compared to those in the control group.

Table 23: Primary analysis

	Unadjusted means				Effect size (adjusted)		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (model)	Hedges g (95% CI)	p-value
Y8 raw score	9396 (1833)	20.9 (20.68, 21.11)	8656 (912)	19.77 (19.54, 19.99)	18052 (17163)	0.037 (-0.07, 0.15)	0.507

Appendix D (Table D1) presents the information used for the calculation of the effect sizes whereas Table 24 presents the model (for raw scores in attainment scales and in logits for mathematics dispositions) based on which these effect sizes are calculated for both the primary and secondary outcomes. For the primary outcome model, we can notice the positive effect of KS2Math (positive), which is very unlikely to have been null, the negative effect of FSM, and some regional variations (when considering the coefficients of certain areas in comparison to the reference, Region 1).

Another observation from the results in this model (as well as those for the secondary outcomes) is that the variance attributed to the school level is quite small, which is also reflected in the small intra cluster correlation: according to this value (0.074), 7.4% of the variation in the total mathematics attainment score can be attributed to differences between schools. For the secondary outcomes this drops further to 6.4% for algebra, 4.8% for multiplicative reasoning, and 2.9% for mathematics disposition.

Table 24: Regression coefficients and ICCs for primary and secondary ITT models

	Maths score		Algebra		Multiplication		Math disposition	
	coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|
Condition (ref: Control)	0.232	0.507	0.119	0.280	0.095	0.496	0.048	0.270
KS2Math	1.200	<0.001	0.300	<0.001	0.540	<0.001	0.019	<0.001
Region (ref: Region 1)								
Region 2	-1.513	0.005	-0.462	0.007	-0.482	0.026	0.074	0.278
Region 3	-1.705	0.004	-0.532	0.004	-0.400	0.088	0.108	0.148
Region 4	-0.719	0.221	-0.174	0.348	-0.285	0.224	-0.010	0.894
Region 5	-0.571	0.302	-0.195	0.263	-0.119	0.587	-0.022	0.742
EverFSM_all (ref: no)	-1.415	<0.001	-0.439	<0.001	-0.543	<0.001	-0.064	0.005
Maths Disp1							0.397	<0.001
Constant	-101.258	<0.001	-25.304	<0.001	-46.010	<0.001	-1.970	<0.001
School ICC (CI)	0.074 (0.06, 0.1)		0.064 (0.05, 0.08)		0.049 (0.04, 0.05)		0.029 (0.02, 0.04)	
N (students)	17,163		17,163		17,163		14,299	
N (Schools)	105		105		105		103	
Wald X^2(p)	30493.86	<0.001	16922.71	<0.001	26629.91	<0.001	4191.03	<0.001
Log likelihood	-55313.74		-36660.6		-42719.6		-21983.9	

It should be noted that the above results do not take into account any changes that took place during the years of the study (therefore students are allocated according to their initial school, even if they changed to the opposite arm). This was to ensure this primary analysis was conforming to the principles of the intention to treat. However, we also ran the same models considering the condition of the school at the end of the study and these results are almost identical (see Appendix 17).

Secondary analysis

Secondary outcomes of interest are the subscales for multiplication and algebra attainment as well as the students' mathematics dispositions. According to the results presented in Table 24 and the effect sizes in Table 25 (see also information in Appendix D), any improvement in either attainment or attitudes resulting from ICCAMS intervention is very likely to be null. The effect sizes indicate the equivalent of one month's progress for algebra and zero progress in multiplication and mathematics dispositions (and the confidence intervals always span zero) when accounting for baseline measures and other variables relevant to randomisation.

Table 25: Secondary analysis

	Unadjusted means				Effect size (adjusted)		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (model)	Hedges G (95% CI)	p-value
Multiplicative score	9396 (1833)	9.095 (8.998, 9.193)	8656 (912)	8.624 (8.52, 8.73)	18052 (17163)	0.032 (-0.06, 0.124)	0.496
Algebra score	9396 (1833)	5.23 (5.17, 5.3)	8656 (912)	4.88 (4.82, 4.95)	18052 (17163)	0.057 (-0.05, 0.16)	0.28
Maths disposition	9356 (1873)	0.22 0.2, 0.25)	8620 (948)	0.2 (0.17, 0.22)	17976 (14299)	0.042 (-0.028, 0.101)	0.27

Analysis in the presence of non-compliance

This section focuses on various analyses performed in order to determine how non-compliance might implicate the results. First, we report on the analysis in response to RQ5 (Is there an interaction between fidelity and attainment change for the treatment schools?) the results of which are presented in Table 26, for both the primary and secondary outcomes.

Table 26: Analysis of primary and secondary outcomes considering the interaction of fidelity and attainment

	Maths score		Algebra		Multiplication		Maths disposition	
	Coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|
Region (ref: Region 1)								
Region 2	-2.11	<0.001	-0.65	<0.001	-0.70	0.01	0.04	0.73
Region 3	-1.59	0.02	-0.43	0.05	-0.36	0.21	0.14	0.21
Region 4	-1.04	0.12	-0.22	0.31	-0.41	0.15	-0.01	0.95
Region 5	-1.07	0.08	-0.32	0.10	-0.24	0.35	-0.06	0.53
EverFSM (ref: no)	-1.39	<0.001	-0.43	<0.001	-0.48	<0.001	-0.07	0.04
Fidelity Score	0.39	0.44	-0.03	0.84	0.09	0.70	0.10	0.37
KS2Math	1.21	<0.001	0.30	<0.001	0.55	<0.001	0.02	<0.001
FidelityXKS2Math	**0.00**	**0.40**	**0.00**	**0.84**	**0.00**	**0.69**	**0.00**	**0.38**
Maths Disp1							0.40	<0.001
Constant	-101.65	<0.001	-25.36	<0.001	-46.79	<0.001	-1.59	<0.001
School ICC (CI)	0.04 (0.03, 0.06)		0.04 (0.02, 0.06)		0.03 (0.02, 0.05)		0.03 (0.02, 0.05)	
N (students)	8,700		8,700		8,700		7399	
N (Schools)	60		60		60		58	
Wald chi square (p)	16120.56	<0.001	8891.09	<0.001	13917.91	<0.001	2101.06	<0.001
Log likelihood	-27842.477		-18430.241		-21581.331		-11323.58	

As shown above, the interaction with the measure of fidelity is negligible for both the primary and secondary outcomes: in other words, taking account of the measure of fidelity makes no real difference to the results and the effect of prior attainment is consistent independently of the degree of fidelity of the intervention at each school.

During the evaluation, it was observed that some schools dropped out and during analysis (data matching with the NPD) it was further noticed that some students changed schools (either from intervention to control, control to intervention, or to other schools within the same arm). From these observations and based on school-level fidelity measurement as discussed previously, we consider it useful to focus on the following non-compliance sources:

- schools dropping out of the intervention;
- students moving in and out of schools (from intervention to control and vice versa); and
- schools within the ‘low’ fidelity category.

These defined three compliance variants, which were then used within 2SLS models as detailed in Appendix 15B. The results from the 2SLS models considering compliance as described above conform overall with the results of the ITT models and there are no significant compliance effects (see Table 15B2 for an overview of the second stage for the primary outcomes and the rest of Appendix 15B for the secondary outcomes and from both stages of the modelling).

A sensitivity analysis performed with relaxing the ITT assumption by removing from analysis the intervention schools that dropped out during the course of the trial (N students = 1,205, and 75 with two schools) was also performed. As shown with the results in Appendix 15C, the effect of the intervention on additional progress in mathematics, algebra, multiplication, and mathematics disposition does not change in any significant way as compared to the ITT results

Further analyses were performed to account for various levels of fidelity for the intervention schools as well as controlling for the effect of the type of teacher as allocated at randomisation. These are presented in additional analyses and further discussed in the IPE section along with other evidence.

Missing data analysis

As already mentioned, the definition of missingness has been a challenging task for this analysis mainly because of the complexities with defining an exact randomisation student sample size. Despite this challenge, we have reported earlier the attrition rates and explore here how missingness at test (DP2) is related to other variables available at student level and based on the initial possible responses our baseline sample.

The models in Appendix 15D present the results of a multilevel logistic regression of missingness at test. They show how the odds of completing a test are reduced with increased KS2 maths score, for girls (there is a high chance these effects were null though) and for lead teachers independently of the condition (intervention or control when compliance was not present in the models). There is also lower chance of missingness with students with higher mathematics disposition at Year 7 and chances are increased for students in Region 3 compared to the reference (Region 1), and also those who have had FSM. These variables were considered in multiple imputation models (Appendix 15E) to explore further the impact of missingness on the outcomes of interest.

Table 27 presents the results for the primary and secondary outcomes using the ten imputed datasets.

The effect sizes involved when accounting for the MI are slightly larger than those from the initial ITT models (also shown for easier comparison on Table 27) and these correspond to about one month’s progress for all (primary and secondary) outcomes but the confidence intervals still cross zero suggesting a very high chance this was null.

In sum, the results in this section show that accounting for missing data marginally strengthens the measures of effect of the intervention but not strongly enough to avoid including zero in the credibility range. We conclude that missing data might be worth following up in future research in general (that is, what is happening when students ‘go missing’ and how does this reflect a phenomenon of learning disengagement more widely?).

Table 27: Results of imputation—ten imputations, 20,198 cases from 106 schools

	Maths score		Algebra		Multiplication		Math disposition	
	Coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|	coefficient	P>|z|
Condition (Ref: Control)	0.551	0.152	0.157	0.165	0.286	0.092	0.070	0.148
KS2Math	1.001	<0.001	0.250	<0.001	0.450	<0.001	0.015	<0.001
Region (Ref: Region 1)								
Region 2	-1.478	0.011	-0.414	0.018	-0.518	0.042	0.054	0.485
Region 3	-1.404	0.026	-0.445	0.017	-0.295	0.287	0.064	0.461
Region 4	-0.561	0.369	-0.124	0.502	-0.237	0.386	-0.026	0.746
Region 5	-0.322	0.583	-0.115	0.511	-0.041	0.872	-0.031	0.679
EverFSM_all	-1.774	<0.001	-0.533	<0.001	-0.696	<0.001	-0.067	0.002
Maths Disp1							0.397	<0.001
Constant	-80.827	<0.001	-20.207	<0.001	-36.782	<0.001	-1.624	<0.001
ES (CI)	0.073 (-0.03, 0.17)		0.07 (-0.03, 0.16)		0.08 (-0.01, 0.17)		0.06 (-0.02, 0.14)	
ES (CI) - ITT	0.037 (-0.072, 0.147)		0.032 (-0.06, 0.124)		0.057 (-0.05, 0.16)		0.042 (-0.028, 0.101)	

Subgroup analyses

Analysis was performed with the subsample of FSM-eligible students and the results are presented in Table 28.

Table 28: Subgroup analysis

	Unadjusted means				Effect size		
	Intervention group		Control group				
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n	Hedges g (95% CI)	p-value
Maths test score	2513	16.79 (16.41, 17.18)	2468	16.49 (16.09, 16.88)	Model N = 4783 Null: 4981	0.063 (-0.037, 0.162)	0.215
Multiplicative score	2513	7.38 (7.2, 7.56)	2468	7.28 (7.09, 7.46)	Model N = 4783 Null: 4981	0.072 (-0.015, 0.159)	0.103
Algebra score	2513	4.14 (4.02, 4.25)	2468	3.97 (3.86, 4.09)	Model N = 4783 Null: 4981	0.076 (-0.032, 0.184)	0.167
Maths disposition (logit)	2499	0.18 (0.13, 0.24)	2452	0.11 (0.06, 0.17)	Model N = 4783 Null: 4951	0.068 (-0.023, 0.159)	0.143

As can be seen, there are slightly larger effect sizes than the models with the full sample, and in all cases (primary and secondary outcomes) these correspond to about one month's progress of the intervention students compared to those in the control condition after we account for baseline attainment and other variables as specified with model (equation 6); however, in all cases there is high chance that these effects were null.

For these statistics we only have missing data at full variable level (that is, from the whole sample analysed of 21,055 there were 6,405 cases with reported FSMever. The models from which the above effect sizes are derived are presented in Appendix 20A and the values used for the calculation of the effect sizes are in Appendix D.

As a further validation check we have also checked for the interactions between EverFMS and condition with the whole sample: the models for primary and secondary outcomes are presented in Appendix 20B. None of the interactions are significant in predicting the outcomes.

Additional analyses and robustness checks

In this section we summarise the results of the additional models run as described in Methods (Table 13) and presented in Appendix 18 (see Tables 29 and 30). It should be noted that there have been some extra specifications of these models which we thought were important to account for after considering the importance and difference in experiences between the lead and cascade teachers. Therefore, we have also run models with this variable (that is, 'teacher type' with reference category 'cascade', as also noted in Table 13), which was available for both control and intervention schools as schools had to declare at recruitment two lead teachers (as part of the recruiting requirements, see School Agreement).

In overviewing the findings from these models (and the multiple outcomes for each tabulation) we first note that in none of these did we find evidence of significant improvement due to the intervention as supported by the coefficient of the condition variable (0 = control/comparator, 1 = ICCAMS). We therefore do not discuss this further but focus on other interesting findings from the results of these models.

The first relates to gender and age, with the results presented in Table 18A (Appendix 18). Girls appear to perform better at the overall mathematics test and the algebra subscale, however they also appear to perform worse than boys in multiplicative reasoning and, as expected from the literature, in mathematics disposition. Age also appears to have a negative, but very small, effect on these measures, which is also evident for algebra and mathematics dispositions where the chance of these being null is very small.

When we consider maths disposition at the start of Year 7, on these models it seems that the gender effect on multiplicative reasoning is gone as well as age effect for the algebra subscale (Models in Table 18B). Having a lead teacher is also associated with higher algebra scores (compared to cascade)—but no other outcome when we also consider maths disposition and teacher type (Table 18C).

Moving to models without disposition but considering transmissionist teaching from the students' perspective (Tables 18D, 18E, and 18F in Appendix 18) we can note the consistent negative effect of transmissionist teaching, even after controlling for teacher type. Teacher type only seems to affect the algebra subscale and the mathematics disposition when also considering transmissionism at Year 7 (Table 18G, Models M3). The message from these models is the consistent association of students' perception of transmissionist teaching on the outcomes, which is very unlikely to have been null. Further models with students' perceptions of transmissionist teaching at the start of Year 7 (Table 18.G, Models M3b) show a positive relationship of Year 7 perception and a negative of their Year 8 teaching perceptions.

The models discussed so far are only based on students' perceptions of teaching practice and other predictors that define the trial (condition, region, and so forth). We now turn to including how teachers perceive their teaching practice within three-level models (see also Models M4a in Tables 29 and 30): we run both for class as well as teacher level (Table 18H shows the teacher-level model whereas more results are presented in Tables 18I and 18J for class level—instead of teacher—with and without teacher type). Even with the three-level models, we observed the consistent effect of students' perception of transmissionist teaching on the learning outcomes whereas there were no effects observed for FA teacher practice in any of the outcomes models.

A final note on the three-level models: there is a noticeable increase in the variance explained by class or teacher level for the attainment outcomes (ranging from 12% for teacher level on multiplicative reasoning to 23% for class level on overall test score). The intra cluster correlation for these levels (class/teacher) remains low (as with school level) for mathematics disposition.

Table 29: Summary of results from various models for attainment outcomes

Maths Score	2-Level Models							3-Level (M4a)	
Covariates	M1	M1a	M2	M2a	M3	M3a	M3b	Teacher	Class
Condition	+	+	+	+	+	+	+	-0	-0
KS2Math	+**	+**	+**	+**	+**	+**	+**	+**	+**
Region	*	*	*	*	*	*	*	ns	ns
EverFSM	-**	-**	-**	-**	-**	-**	-**	-**	-**
Gender	+**	+**	+**	+**	+**	+**	+**	+*	+*
Age	-	-	-	-	-	-	-	-*	-*
TeacherType		+		+		+	+	-	-
Maths Disposition1			+**	+**					
TransTeaching2					-**	-**	-**	-**	-**
TransTeaching1							+**		
FAPractice(T1Y8)								-	-
TransTeach(T1Y8)								-*	-
Algebra	2-Level Models							3-Level (M4a)	
Covariates	M1	M1a	M2	M2a	M3	M3a	M3b	Teacher	Class
Condition	+	+	+	+	+	+	+	+	+
KS2Math	+**	+**	+**	+**	+**	+**	+**	+**	+**
Region	*	*	*	*	*	*	*	ns	ns
EverFSM	-**	-**	-**	-**	-**	-**	-**	-**	-**
Gender	+**	+**	+**	+**	+**	+**	+**	+*	+*
Age	-*	-	-*	-*	-*	-*	-*	-*	-*
TeacherType		+*		+*		+*	+*	+	+
Maths Disposition1			+**	+**					
TransTeaching2					-**	-**	-**	-**	-**
TransTeaching1							+*		
FAPractice(T1Y8)								-	-
TransTeach(T1Y8)								-*	-*
Multiplication	2-Level Models							3-Level (M4a)	
Covariates	M1	M1a	M2	M2a	M3	M3a	M3b	Teacher	Class
Condition	+	+	+	+	+	+	0	-	-
KS2Math	+**	+**	+**	+**	+**	+**	+**	+**	+**
Region	*	ns	*	ns	*	ns	ns	ns	ns
EverFSM	-*	-**	-**	-**	-**	-**	-**	-**	-**
Gender	-*	-*	-	-	-*	-*	-*	-*	-*
Age	-*	0	0	0	-0	0	+0	-	-
TeacherType		-0		-		-0	-	-	-
Maths Disposition1			+**	+**					
TransTeaching2					-**	-**	-**	-*	-*
TransTeaching1							+**		
FAPractice(T1Y8)								-	-
TransTeach(T1Y8)								-*	-

Notes
-/+: indicate the sign of coefficients (when relevant)
n.s: not significant (for categorical variable, otherwise non significance implied with no star)
* p <0.05, **p <0.001
0 – coefficient approaches 0

Table 30: Summary of results from models for maths disposition

Maths Disposition 2	2-Level Models					3-Level (M4a)	
Covariates	M1	M1a	M3	M3a	M3b	Teacher	Class
Condition	+	+	-	-	-	-	-
KS2Math	+**	+**	+**	+**	+**	+**	+**
Region	ns	ns	ns	ns	ns	ns	ns
EverFSM	-*	-*	-*	-**	-*	-*	-*
Gender	-**	-**	-**	-**	-**	-**	-**
Age	-*	-*	-*	-*	-*	-	-
TeacherType		+		+	+	0	0
Maths Disposition1	+**	+**	+**	+**	+**	+**	+**
TransTeaching2			-**	-**	-**	-**	-**
TransTeaching1					+**		
FAPractice(T1Y8)						+	+
TransTeach(T1Y8)						+*	+*

Analysis with only intervention schools

Models were run only with the intervention school sample to also explore the confidence of teachers to teach ICCAMs as well as the school fidelity. The three-level model (teacher level) for the primary outcome is presented with Table 31.

Table 31: Three-level (student-teacher-school) model of primary outcome including school fidelity

	Coefficient	Standard error	z	P>|z|
KS2Math	1.028	0.019	53.2	<0.001
Fidelity Score	0.006	0.171	0.03	0.973
Region (ref: Region 1)				
Region 2	-1.665	1.098	-1.52	0.129
Region 3	-0.966	1.166	-0.83	0.407
Region 4	-1.25	1.113	-1.12	0.261
Region 5	-0.189	1.043	-0.18	0.856
EverFSM	-1.215	0.226	-5.37	<0.001
Teacher type (ref: cascade)	-0.295	0.308	-0.96	0.338
Gender (ref: Boys)	0.307	0.19	1.61	0.106
Age (in months)	-0.057	0.027	-2.13	0.034
TransmTechingDP2 (students)	-0.147	0.16	-0.92	0.358
TeacherFA practice@DP2	0.086	0.527	0.16	0.87
Teacher TransTeaching@DP2	-0.674	0.642	-1.05	0.293
Teacher ICCAMS confidence	0.193	0.147	1.31	0.189
Constant	-75.035	4.734	-15.85	<0.001

Models further included an interaction of the fidelity score with teacher type, and also with KS2 maths score, and were run for secondary outcomes as well (see Appendix 19): there was no evidence of additional impact of any of the other variables on maths test results apart from FSM and age. The same patterns are observed when looking at algebra. The model for maths disposition at DP2 (and algebra subscale to a smaller degree) confirms previous findings of the effect of students' perception of transmissionist teaching (and gender). When adding fidelity in the models, we have not observed any impact on any of the outcomes. Please note that models with interactions have already been presented under the section on non-compliance.

Modelling change in perceptions of teaching practices

Given the significant drop of numbers in the analytical sample when we consider teacher self-reports of teaching practices, to cross-examine evidence from other elements of the IPE, it was deemed useful to explore the impact of the intervention on (changing) students' perceptions of transmissionist teaching they experience. The outcome modelled in this case is students' perceptions of transmissionist teaching practice at the end of Year 8 accounting for their perceptions at the start of Year 7 (for transmissionist teaching), the teacher type, and other variables which are included in the ITT models. We had further included in this explanatory model another available relevant variable, that of students' perception of the difficulty of the mathematics lessons (an ordinal variable with reference category 'too easy').

Model 1 (in Table 32) only includes key variables in the main models and explores the effect of conditions and teaching style in the change of students' perception of transmissionist teaching.

Table 32: Modelling of students' perceptions of transmissionist teaching at Year 8

	Model 1		Model 2		Model 3		Model 4		Model 5	
	Coef.	P>|z|	Coef.	P>|z|	Coef.	P>|z|	Coef.	P>|z|	Coef.	P>|z|
TransTeaching@ Year 7	0.29	<0.001	0.29	<0.001	0.29	<0.001	0.26	<0.001	0.24	<0.001
Condition (ref: control)	-0.12	<0.001	-0.09	0.02	-0.09	0.02	-0.09	0.02	-0.09	0.01
Teachertype (ref: cascade)	0.00	0.97	0.04	0.09	0.04	0.10	0.03	0.16	0.02	0.28
Condition#Teachertype (ref: Control/cascade)			-0.07	0.02	-0.07	0.02	-0.07	0.03	-0.05	0.08
Newtotalscore					-0.01	<0.001	0.00	<0.001	0.00	0.15
KS2Math	0.00	0.24	0.00	0.21	0.01	<0.001	0.01	<0.001	0.01	<0.001
Region (ref: Region 1)										
Region 2	-0.03	0.63	-0.02	0.66	-0.02	0.69	-0.01	0.90	0.00	0.98
Region 3	-0.11	0.07	-0.11	0.07	-0.11	0.09	-0.10	0.10	-0.06	0.32
Region 4	-0.15	0.01	-0.15	0.01	-0.14	0.02	-0.14	0.02	-0.14	0.01
Region 5	-0.03	0.58	-0.04	0.52	-0.03	0.60	-0.03	0.63	-0.02	0.70
EVERFSM	-0.01	0.51	-0.01	0.51	-0.02	0.17	-0.01	0.42	-0.02	0.06
Ageinmonths					0.00	0.60	0.00	0.38	0.00	0.13
Lesson_difficulty (ref: [too easy])							@Year7		@Year8	
2 [about right]							-0.03	0.05	-0.22	<0.001
3 [too hard]							-0.04	0.28	0.02	0.46
MathsDisposition@Year7/8							-0.04	<0.001	-0.16	<0.001
Gender					0.02	0.09	0.01	0.37	-0.02	0.08
Constant	0.20	0.07	0.18	0.11	-0.27	0.27	-0.16	0.54	0.14	0.55
N (students)	13255		13255		13122		11807		12030	
N (Schools)	93		93		92		92		92	
N (Teachers)	557		557		552		551		551	
Wald chi2(16)	1001.08		1006.94		1039.87		1008.88		2820.91	
Log likelihood	-12154.9		-12152.06		-12017.28		-10787.97		-10186.39	

There is a consistent effect of the condition (dummy for intervention—so the negative effect indicates that intervention students perceive their teaching as less transmissionist) throughout the models, even after controlling for other variables such as perceived difficulty of mathematics lessons, score on test, and so on.

The apparent non significance of teacher type (effect reported above) is most likely the artefact of a hidden interaction effect as shown with Model 2: it is initially unlikely to be null but then after adding more variables to the model (disposition and perception of lesson difficulty, Models 4 and 5) it is not. What this means is that the effect of lead teachers (compared to cascade teachers) across the whole sample is mediated by their students' (more positive) dispositions to maths and their lower perception of the difficulty of lessons. The causality involved cannot be ultimately determined from these associations but we hypothesise this is related to the lead teachers being more positively engaged with, and engaging of, their students (alternative hypotheses are that they tend to be chosen to teach the more positively disposed students and classes, or a combination of the aforementioned).

To sum up these additional investigations up to this point, while the intervention has in almost all models indicated non-significant improvements over the comparator group, there have been some significant mediators, and some of those are related to teachers' practice and learners' perceptions of this practice: these have proved to be significant in favour of the intervention schools and robust in the models controlling for many salient variables. If this is substantiated, it would indicate that the intervention has had measurable, significant effects, but at the level of the teacher, teaching practice, and learners' perception of the teaching practice and not at the level of students' performance intended to measure learning outcomes.

We now turn to the IPE to explore these findings in more depth, also to situate them in the context of the implementation of ICCAMS in practice.

Implementation and process evaluation results

The process evaluation was designed to explore how the ICCAMS intervention was implemented in schools and its effect on teachers' and learners' practices and outcomes. Findings are based on a combination of evidence from fieldwork with a sample of eight of the 49 intervention schools that completed the programme, from survey data from teachers (see Table 35) and students from participating schools, and from an additional school-level survey of the control schools. This part of the evaluation focuses on the professional development, the fidelity with which the schools conducted the programme, and teacher and student attitudes to the project. This leads to a detailed discussion of credible explanations, or causes, of outcomes and hence of recommendations for the future. This process evaluation identifies likely strengths and weaknesses in the intervention compared to the control practices and suggests possible reasons why the intervention might have had a stronger or weaker impact on teaching, learners' attainment, and other outcomes compared to control. One of the key findings from this part of the evaluation was that the intervention was operationalised differently in the different schools, with some paying closer attention to, and engaging with, the programme intentions more than others. We also explore why this might have occurred. The focus here is on what happened to the PD inside the school and its effect on the classrooms. Additionally, a key limitation is identified: we found a need to know more, indeed a lot more, about what was occurring in the control schools (as also described in the school participation agreements).

We opted for structuring this part of the report under two main analytical approaches. In the first part we amalgamate and triangulate evidence from the various elements of the evaluation around the key elements of the programme adopted (informed by the suggested themes of the EEF's reporting template); we call this a 'thematic descriptive analysis of implementation, process, and fidelity'. The second part is more reflective and theoretically informed and focuses on establishing possible explanations of the evidence through a single case study of the 'impact of ICCAMS on learning and teaching outcomes'.

An important note about both sections is that we put forward most of this conceptual and empirical analysis (that is, most of the thematic analysis and the case study) before the statistical analysis of the outcomes reported earlier: we believe this strengthens the validity of our analytical results and helps give credence to our explanations, together with what we will refer to as mediating and moderating conditions for these explanatory causes.

Professional development days in all five regions were attended by research staff and we interviewed the training leads. Interviews with teachers were conducted and lessons observed. Teacher surveys were conducted at the start and end of the project and the sample sizes for this analysis are presented in Table 33. Here 'lead/cascade teachers' in the control schools refer to those nominated as such prior to the randomisation and do not imply they were actually engaged in leading activities, though they may well have been, on other projects. These responses represent a relatively high proportion of teachers in the participating schools based on the available information: this drops from 83% and 92% at DP1 to 54% and 63% at DP2 for cascade teachers at intervention and control schools, respectively; for lead teachers there is an observed increase in participation from around 77% at DP1 to 82% at DP2, with 90% at intervention and 71% in control schools at DP2. (See Appendix 22 for sample descriptors by region.)

Table 33: Teacher sample sizes by teacher type, condition, and data point

	DP1 (Autumn 2016)			DP2 (Summer 2018)			Total
Teacher Type	Intervention	Control	Total DP1	Intervention	Control	Total DP2	
Cascade T	211 (252)	227 (247)	438 (499)	138 (253)	152 (240)	290 (493)	728
Lead T	84 (108)	64 (84)	148(192)	66 (73)	40 (56)	106 (129)	254
Missing (no information on teacher type)	1	9	10	0 (11)	28 (11)	28 (22)	38
Total	296 (360)	300 (331)	596 (691)	204 (337)	220 (307)	424 (644)	1020

Note: the bracketed numbers next to the sample sizes present the possible number of teachers identified in the school lists shared by the schools (for the students in the corresponding year groups at every year of the study). It should be noted, though, that this information was available from 100 schools at DP1 and 97 schools at DP2 out of the 105 that provided pupil data.

To facilitate interpretation of the findings below the reader is reminded of the ranges of the measures mentioned throughout via the boxplots in Figure 12. These are mainly presented for descriptive purposes at this stage and are not intended for making causal inferences.

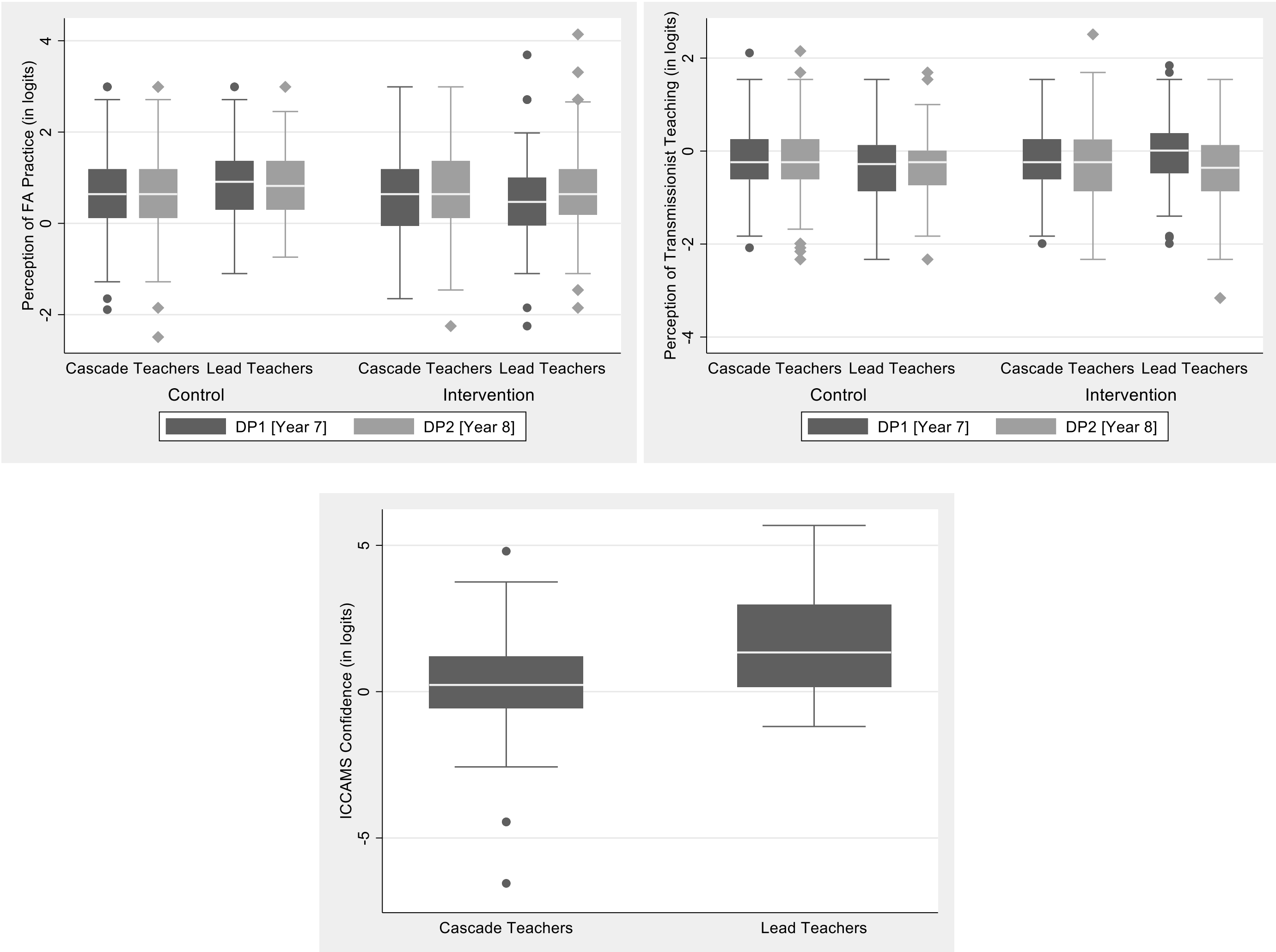


Figure 12: Boxplots of available teacher responses on perceptions of FA practice and transmissionist teaching (top figures, also showing intervention vs comparator group) and ICCAMS confidence (bottom, only intervention teachers) by teacher type

The shaded parts of the boxes show the central 50% of the distribution (divided by a line representing the median measure) of the teachers' self-report scores for the three outcome measures 'transmissionism', 'formative assessment practice', and 'confidence in teaching ICCAMS' respectively. In particular, the top two figures show the distribution of teachers' measures on perceptions of 'FA practice' and 'transmissionist teaching' at DP1 (that is, teaching Year 7) and at DP2 (teaching Year 8). The bottom figure presents the distribution of the measure of 'ICCAMS confidence', which was only captured in intervention schools at DP2 (end of Year 8). The graphs therefore imply (1) noticeable reductions in transmissionism for lead teachers between DP1 (Year 7) and DP2 (Year 8), (2) some small increases in FA practices between DP1 and DP2, and (3) substantially higher confidence of lead teachers than cascade teachers in teaching ICCAMS overall. These results are further discussed at subsequent relevant sections in the report.

In order to help readers with interpretations such as this in the following sections, we will insert a qualitative descriptor where this seems important according to whether the value is above the upper quartile (**High**), within the interquartile range (**Medium**), or below the lower quartile (**Low**). (Details of the definitions of these cut-offs are provided at the end of Appendix 14.)

Descriptive thematic analysis of IPE evidence

This part of the IPE will present thematic evidence on how the intervention worked. Doing so involved considering the conditions that are regarded as necessary for its success in schools and exploring the relevant evidence. These conditions as we understand them from the design of the developers, after the lead teachers had attended the regional PD sessions, were as follows:

- lead teachers implement the programme in their own classrooms and schools;
- there is time and support for the cascade PD to replicate the key processes required for the cascade teachers;
- teachers understand and engage with ICCAMS intentions and designed practices;
- teachers have good relationships with students and behavioural management skills and so forth in order to implement these in classroom practice;
- departments are stable and open to new approaches; and
- support from school management enables all the above.

These conditions provide a bit more nuance to understand the complexity of such interventions, which goes beyond the simplistic measure of school fidelity described above. We explore next how they were implemented under six main themes: professional development, implementation of ICCAMS in schools, fidelity, improvement recommendations from the stakeholders' perspective, and business-as-usual practice.

Professional development in practice

We focus here on RQ6 (a and c) and at some points on RQ7 (cascading) to explore the evidence in relation to training the trainers to lead the PD as well as the lead teachers to teach with ICCAMS.

Training the trainers—preparing and supporting professional development leads

The developer and delivery teams' work with the five regional professional development leads took place on six full day meetings over the two year period. In these meetings the ICCAMS originators and developers met with the PD lead, discussing provisions for each of the regional PD sessions and their aims in helping the teachers both to implement ICCAMS teaching in their classrooms and to cascade their training to other teachers in their schools. In these meetings, the PD leads were able to consider the mathematics and form a way of working with the teachers.

They would often spend time in pairs going through a lesson, looking at the background mathematics, and considering the kind of responses they would get from teachers and pupils. Then they would present that to others in the group, with the developers commenting, which the PD leads said they found useful because it helped inform their understanding of the materials and how the ICCAMS principles were being exemplified—this helped to prepare PD leads to cascade the training to their colleagues.

With regard to the materials, one PD lead said:

> *'In terms of the resources required to run the training days, we had everything we needed: handouts, PTTs, the lot. They were quite detailed notes about how to run the training and we always discussed them in groups' (PD lead, Region 5).*

They also mentioned how the PD leads worked together after each PD session they had conducted, feeding back and reviewing the training:

> *'So after each session, if there was anything worthy of note, we used to just put on an email to everyone, saying so "this is what happened, this went really well, this probably needs tweaking a bit". So it was very useful, so a very strong team, who worked really well together' (PD lead, Region 5).*

One PD lead interviewed referred to the quality of the ways the days ran and noted particularly the 'modelling' that might have informed their own approach to their regional lead teacher sessions:

> *'They were really good. They were quite hard work, I found, in terms of just the intensity of them. I think they were … there was a degree of … fluffiness around them; the agendas were loose, it felt. But also, it felt they modelled, because of, I think they would have tightened it up if we hadn't done what we needed to do. The type of people we are, I think we sat down and engaged with it, and I think that was the model that carried it through' (PD lead, Region 4).*

The same PD lead commented on the significance of making the key mathematical ideas in every lesson explicit:

> *'I think it sharpened up the understanding of what, how the ICCAMS principles were being exemplified. I think it also pulled out, for me, so what are the really key, what is the "big idea" in this thing? Because the lessons are very rich, and I think one of the problems the teachers found at the beginning, and I think this came out through the lesson observations, was, right, at the end of the lessons, what are we pulling together, what have we done, what's the key bit that I want to take away? I think that that was, maybe a challenge for all of us. And so pulling out that content, what was the key bit of content, what was the key idea. And I think that was where those conversations really helped' (PD lead, Region 4).*

A PD lead from a different region made a similar point about the nature of the training:

> *'[The developers] deliberately engineered that so that we experienced those mathematical 'a-ha' moments that we could then engineer into our PD sessions' (PD lead, Region 2).*

Another added how the materials provided for their subsequent regional lead teacher sessions engendered confidence:

> *'For the training we delivered ... we got a very, very detailed PowerPoint presentation and it needed very little adaptation, other than to decide which bits could be missed out. So it was there, kind of there if you needed it. Sometimes you would get the responses from teachers that could eliminate three or four slides. Sometimes you could have less of a prompt. So they were, on the whole, a good quality, so you'd have the order of 60 or 70 slides for a day, each of which you would have to choose; some were obviously critical, setting teachers the task, some you could use them as backup if you needed it, if you weren't getting any response from teachers' (PD lead, Region 5).*

Lead teacher training (including attendance) in the five regions

The professional development programme aimed to involve two teachers from each school, both attending (together) nine training days across two years—six in year one and three in year two. This corresponds to a maximum of 18 teacher days of attendance per school during the programme. Figure 13 presents the attendance figures for schools overall and per region.

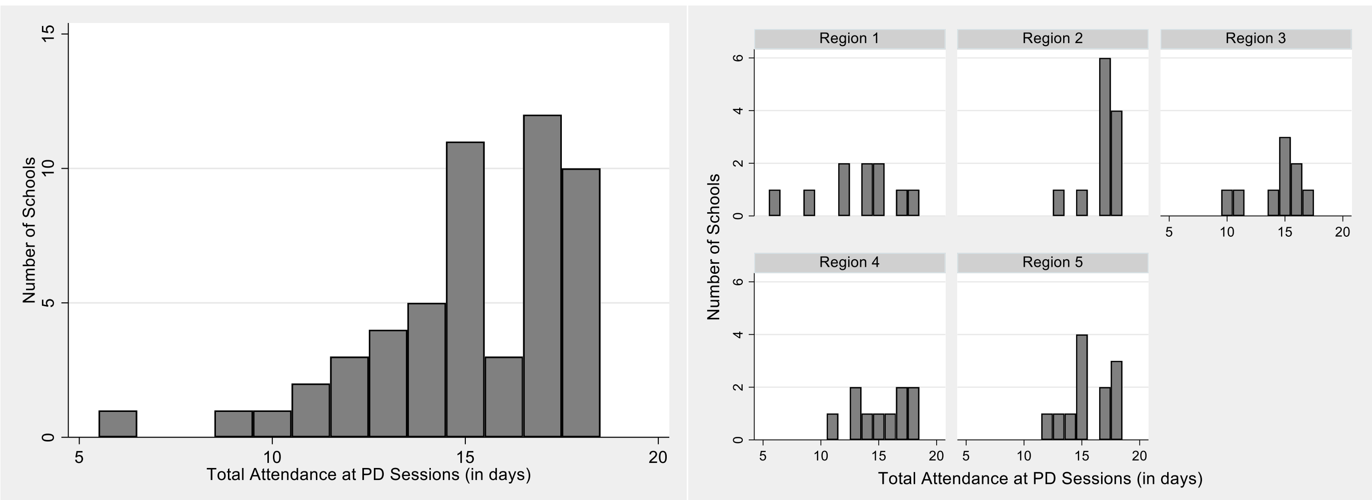


Figure 13: Histograms of attendance on PD—overall and by region, measured in teacher days

Figure 13 shows that, overall, the regional PD sessions were well attended with the majority of schools attending 15 days or more out of a possible 18. There were only a few schools with attendance below ten teacher days. It also shows the particularly high attendance in the Region 2 (a more detailed breakdown by region is presented under the Fidelity subsection).

Each PD lead ran the training in their designated region, instructing the groups in their discussions of past and future lessons, and reviewing student and teacher experiences. In our observations we concluded the principles of the ICCAMS project were discussed in some detail, and some discussions also included the cascade training and its implementation.

From our observations of the training days we judged that the PD for lead teachers was well-received in all regions: teachers were involved, asking questions, talking with each other about concerns and issues—really participating and evidently learning. Region 2 showed particularly high teacher attendance (close to 100% across the two years) and engagement and our observations suggested the engagement was effective for all teachers present. In other regions the attendance was around 75% and their engagement and participation were also good. According to teacher survey responses, overall, lead teachers perceived the training useful and also helpful for their confidence with the majority agreeing or strongly agreeing with the statements in Figure 14 (the lower agreement rates of cascade teachers will be discussed later).

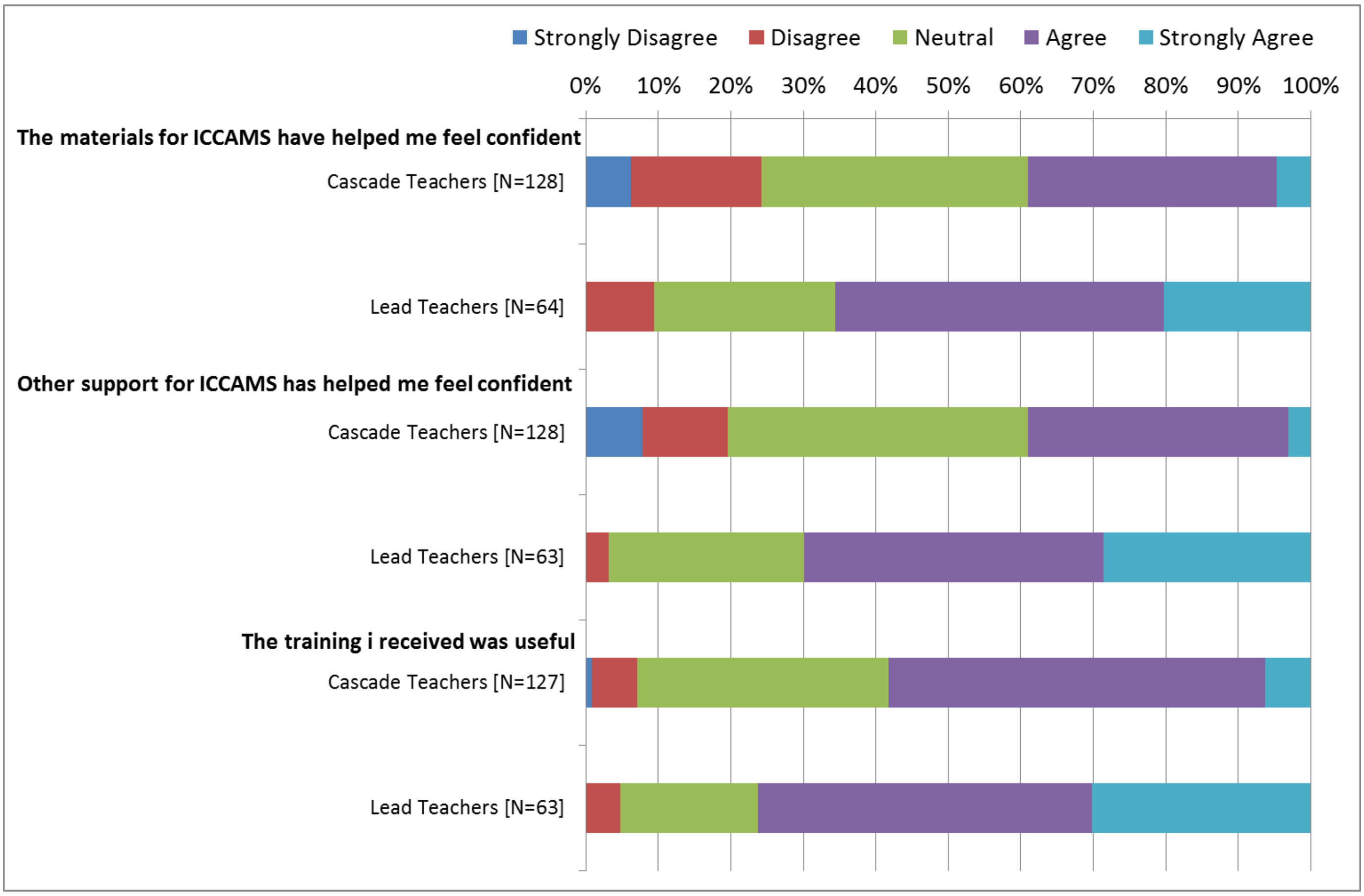


Figure 14: Distribution of teacher responses on various aspects of ICCAMS, by teacher type

The ICCAMS materials and PD supported the lead teachers in their understanding of ICCAMS principles and teaching practice; many discussions focused on particular ICCAMS activities and their classroom implementation, which the teachers found helpful. Our survey further showed that from the 64 lead teachers who answered the question, 47% rated the PD as 'excellent', with 34% reporting 'good'. Their ratings on various elements of the ICCAMS PD are summarised in Figure 15. 'Working through maths problems' was the element found most helpful by teachers during the PD sessions (with 98% reporting this was quite/extremely helpful). 'Reflections on lessons taught' and 'planning for new lessons' were also highly commended (both with 89% 'quite/extremely helpful' ratings) and then also 'exploring examples of children's work' (85%).

At its best, the PD gave lead teachers the time to think about their teaching and the mathematics involved from different perspectives, revealing different methods and approaches. Interview evidence suggested that discussing approaches with lead teachers from other schools was especially appreciated: this reflects the added value from gathering lead teachers from different schools. One lead teacher mentioned, for example:

> *'The main thing is the sharing of ideas and learning. And just the exploring was new and actually talking to other teachers, 'why have you done it that way?' and how has someone else taught it to their class … how they extended them? … how they are putting this sort of thing into their schemes?' (School 84, Lead teacher, female, TY7=Medium, TY8=Low, FA7= Medium, FA8= Medium, IC=High, ID: S084T007).*

The same teacher further expanded on these opportunities for sharing ideas with others, sometimes in ways not experienced since entering the profession:

> *'I've always been one to show the different methods on working things out and letting students pick for themselves what they want to do. I've always been on that route, I've never said, 'this is the only way you can do it', but I think it's opened my eyes to how much I've missed, like, how many other ways I've not even considered. And it might be things that I used to do but just in a seven or eight years of teaching, I've lost it somewhere! And then got into a routine. It's just refreshing isn't it, to remember what you did?! Like I think I explored more when I was on my PGCE, doing things like that because I had more time. And that is the time, to explore, they always tell you. But actually it's not, you explore throughout teaching' (School 84, Lead teacher, female, TY7=Medium, TY8=Low, FA7= Medium, FA8= Medium, IC=High, ID: S084T007).*

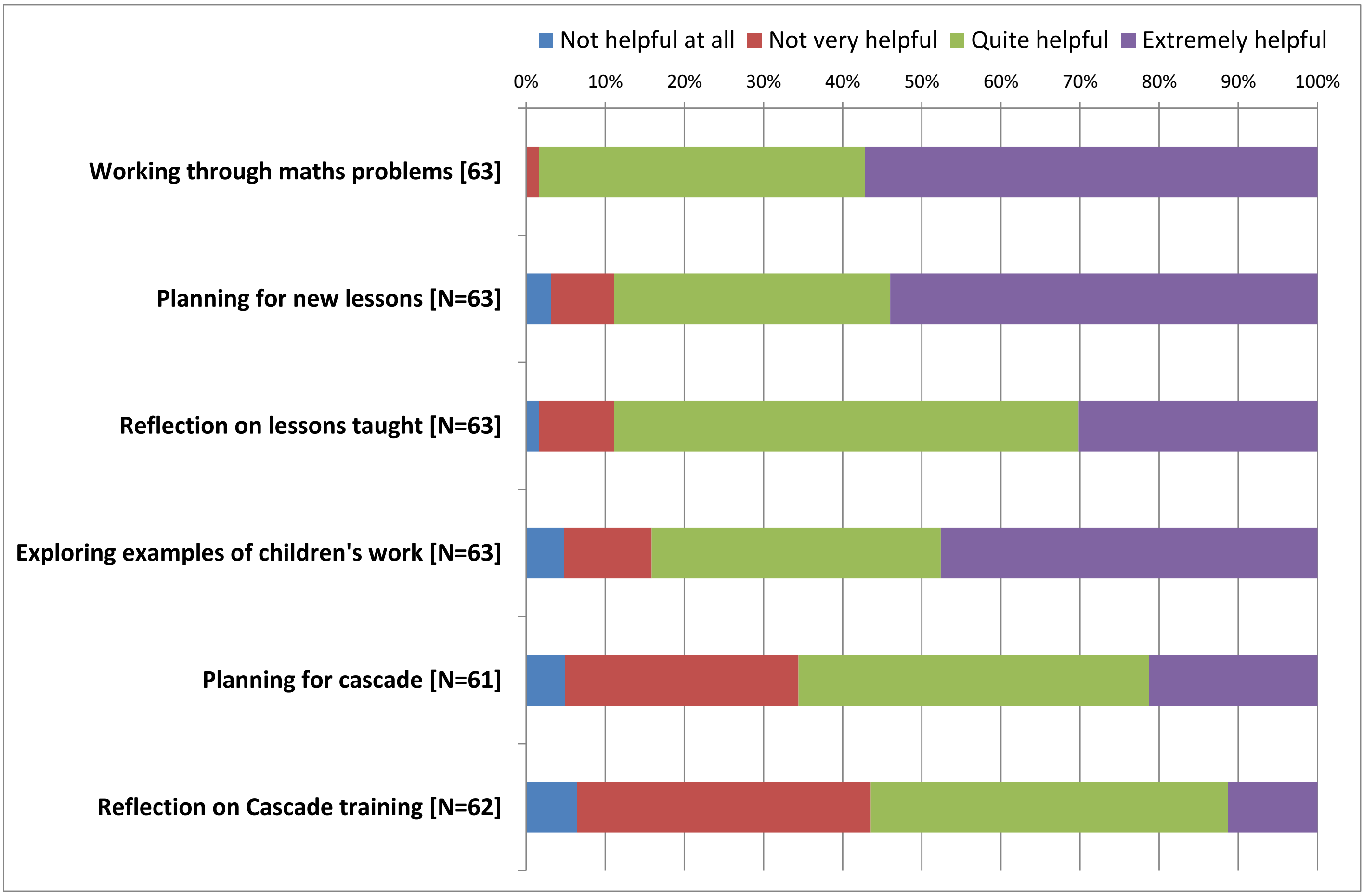


Figure 15: Lead teachers' distribution of responses to question, 'How helpful and informative do you find each of these elements of the ICCAMS professional development?'

Teachers often expressed surprise at the different methods used to tackle problems, statements such as, 'Some things, I am like, "Oh, wow, I've never thought of it in that manner!"' (School 84, Lead teacher, female, S084T007) were not uncommon in interviews and observations. The depth of discussion, looking at the ideas and design of the lessons, seemed to ensure that many lead teachers felt better prepared. Reflecting on the lessons after they had been taught, their comments revealed that students could not always access the mathematics at the level teachers had expected; this also highlighted misconceptions, which was valued by the teachers. The following quote from the PD lead of Region

4 is illuminating in this regard. In the first PD session, the lesson was considered by some teachers to be too easy for their students, but in reality the students found it more challenging than the teachers expected:

> *'Day two [of the lead teachers' PD days], people came and they did not know, they could not do it, … and other people said, "Oh, and I could not either." That really created the buzz, that was the thing that was needed, that was a great lesson to start with, because it was so straightforward, it was so simple, but equally they could not do it, they could not access it at the level teachers thought they could. And there was a surprise' (PD lead, Region 4).*

While lead teachers from each school had these regional days away to interact with the PD leads and the other teachers, the cascade training for other maths teachers in their schools was intended to take one hour, and only conducted nine times (after each of the nine regional PD sessions). In such timescales, the in-school sessions always felt short of time as they were effectively trying to replicate a day's work away from school in one hour in school (with all that involves in terms of staff being drawn away into other activities and priorities, and being conscious of thinking about other immediate school-related tasks). The evidence is that this was reflected in the lead teachers' perceptions of the cascade: they perceived the planning and reflecting on cascade processes in the PD sessions to be much less helpful than their work on the lessons themselves.

Provision of support for lead teachers in developing their cascade training was discussed in the lead teachers' PD sessions, and lead teachers expressed their opinion about this element, as also shown in Figure 15: 37% of teachers found 'reflections on cascade training' as 'not very helpful' whilst 6% rated them as 'not helpful at all'; 30% said the 'planning for cascade training' was 'not very helpful' with 5% saying 'not helpful at all'. These percentages compare unfavourably with those items related to their own teaching of ICCAMS, such as exploring the lessons or problems themselves, or planning and reflecting on their lessons where the combined 'not helpful' responses do not exceed 15%.

This is likely the result of lead teachers' frustrations with providing the cascade training in their schools as much data to come below suggests many reasons why the cascade proved difficult (see next section and Fidelity section). Some of these reasons and possible explanations are further discussed in the conclusion reflecting on relevant literature.

In the summer term of each academic year, the PD trainers visited each school and observed one lead teacher and one cascade teacher on each of the two visits; 88% of lead teachers found their experience of this helpful. The teachers had good relationships with the trainers and understood these observations were non-judgemental with the aim of helping the teachers in their practice.

The ICCAMS Teaching Material

A key element of the ICCAMS intervention in practice, and the PD (including the PD events), is "The ICCAMS Teaching Materials: teacher's handbook". The handbook provides very comprehensive and detailed support for teaching the key lessons and conducting the mini-assessments. A key quality of the materials is that they do not simply provide lesson plans and handouts, but make the link to Formative Assessment practices, especially this is explicit in the provision of the "mini-assessment" material.

Overall the handbook was well received by PD leads, and intervention and cascade teachers. Figure 16 shows teachers' responses on the parts of handbook they refer to when planning ICCAMS lessons.

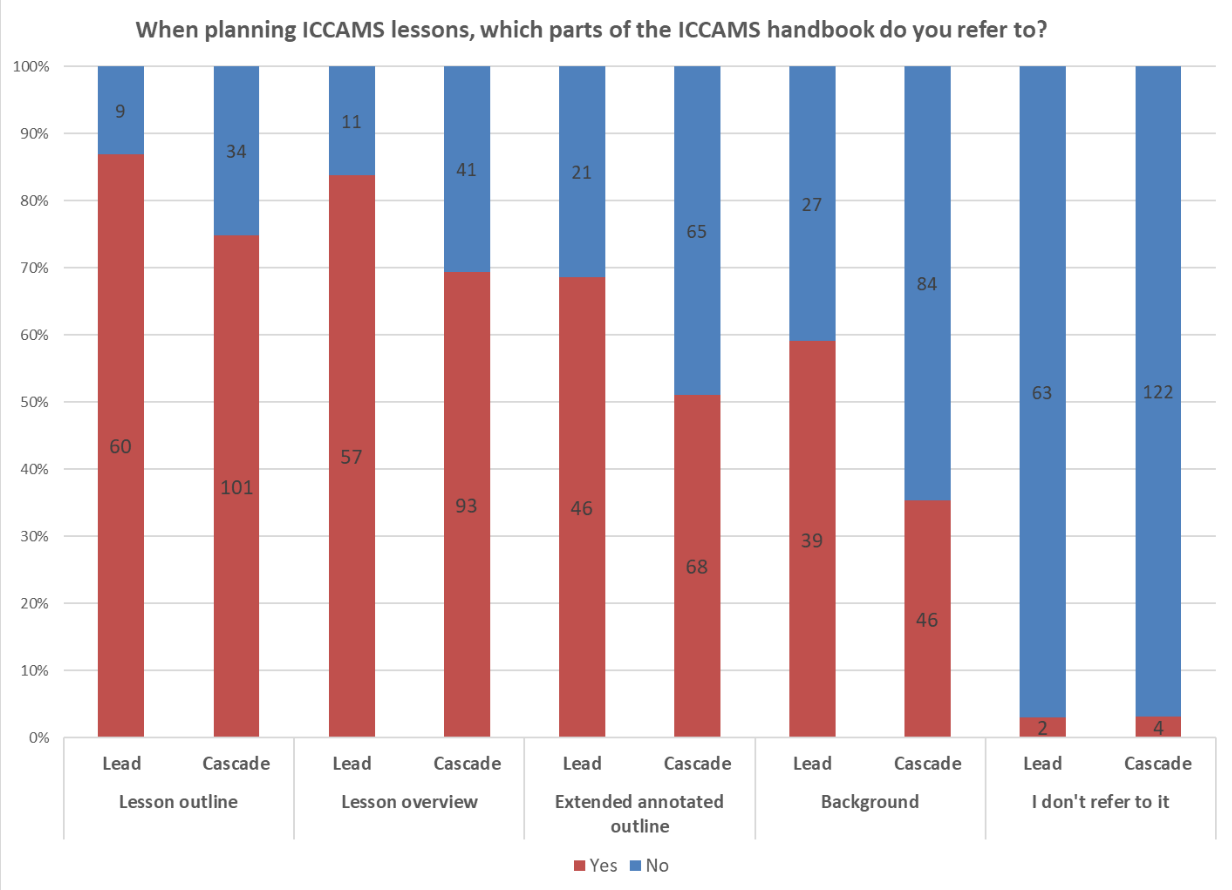


Figure 16: Lead and Cascade Teachers' distribution of responses to question "When planning ICCAMS lessons, which parts of the ICCAMS handbook do you refer to?" [numbers in the bars are available respondents' numbers]

As shown, the most popular part for both lead and cascade teachers are the lesson outline and lesson overview (more than 80% for lead teachers and 70% for cascade). The percentages of references to extended annotated outline and background drop for both groups, with the latter falling below 50% for the cascade teachers.

However, there were some challenges, identified based on combined evidence from observations, teachers' reports and survey responses.

We first look at teacher's views on the materials and approach. An example from teacher interviews is given below, where one lead teacher explains how another may have been challenged with this material – noting also the improvement due to the intervention:

> *We have our traditional textbook teachers. I think they in particular have found this more challenging because they don't have that culture of discussion around it. And I think it's [the intervention] helped a few people with that. (*School 82, Lead teacher, male, ID: S082T004*)*

The topic was also popular in teachers' responses to the open question in the teacher questionnaire (at DP2) "Please describe any significant difficulties / barriers you have encountered when planning or delivering ICCAMS lessons." From the 103 responses by teachers (35 from lead teachers and 68 from cascade teachers), 50 statements were explicitly related to the materials (14 from lead teachers and 36 from cascade teachers), with the rest mainly focusing on staffing, time constraints and implementation with low ability (these are discussed later). The main interconnected themes of teachers comments are listed below along with some illustrative quotes to cover the spread of comments:

Table 34: Overview of key themes from the open responses of teachers on difficulties related to the materials, when planning or delivering ICCAMS lessons

Theme (with example quotes from teacher open responses)	Frequencies(*)	
	LT	CT
Differentiation/Scaffolding/Adapting material • Found them much more difficult to deliver to lower ability students as the numbers chosen in the lessons were often difficult for such students to access (LT) • Helping other members of staff adapt lessons for low ability sets who do not have the prior knowledge. (LT) • Lack of differentiation for mixed-ability classes. (LT) • The resources are not scaffolded enough for lower ability pupils. For them to complete one problem, there needs to be a lot of sign posted steps. The problems often require a level of English comprehension that is difficult for many students. (CT)	10	12
The contents of the lessons/handbook • Not enough structure. Too open ended. (CT) • Some of the lessons can be very repetitive for students (CT) • The resources provided on the website are very basic and include only a couple of slides. Going forward the entire lesson should be available on a PPT file including the extra questions that are found in the book (CT) • the PowerPoint quality became less the further along in the scheme of learning I went. (CT)	3	9
Teaching and learning issues • We often found the ICAAMs lessons undid the good we had done and then confused the students with the methods they already know, e.g. multiplication with arrays. • Some of the methods led to pupils not using the skills that they already have for basic maths, and they sometimes became confused about which methods to use. • Understanding what the plan is getting at. • How to communicate the objective of reasoning and multiple methods		10
Fit in lesson timings • At times I have had to adapt some of the lessons to ensure that we get most of the content covered within one lesson, or added additional questions to extend the lesson to ensure that we don't finish too early, but this is rare. (LT) • Generally not enough content to fill a full hour. The lessons were very unstructured …(CT)	2	4
Lack of engagement from pupils • Pupils did not engage with the tasks. Lessons did not challenge the pupils and always felt like a step backwards from their learning. • The students did not always find the contexts engaging and I found it hard to get them to buy in at times.		4
(*) As many teachers mentioned multiple difficulties the frequencies do not sum up to the total responses mentioned earlier		

In sum, the most frequently reported difficulties encountered by teachers in relation to the materials referred to challenges with differentiation and 'scaffolding' especially with low attainers. There were also comments on the contents and (the lack of) structure of the material which some teachers reported as difficulties, and extra time demands (for example, a few referred to extra time needed to "make the PowerPoints more student/lesson friendly" (LT) and to "make the lessons more effective"). There were also some reported problems of student engagement with the context of the lessons and some teaching and learning issues with some lessons in the form of reported negative impact on students' learning and teachers' understanding of the plans and what they are expected to do for "communicating the objectives" of the lessons, In addition to the themes listed in Table 34, more direction was also explicitly asked for by two CTs: "Answers would be useful to save time", "Some videos of delivery would be nice".

Figure 17 illustrates how closely teachers follow the material, and as shown the majority of teachers report that they made changes, and this was also more prominent within the cascade group.

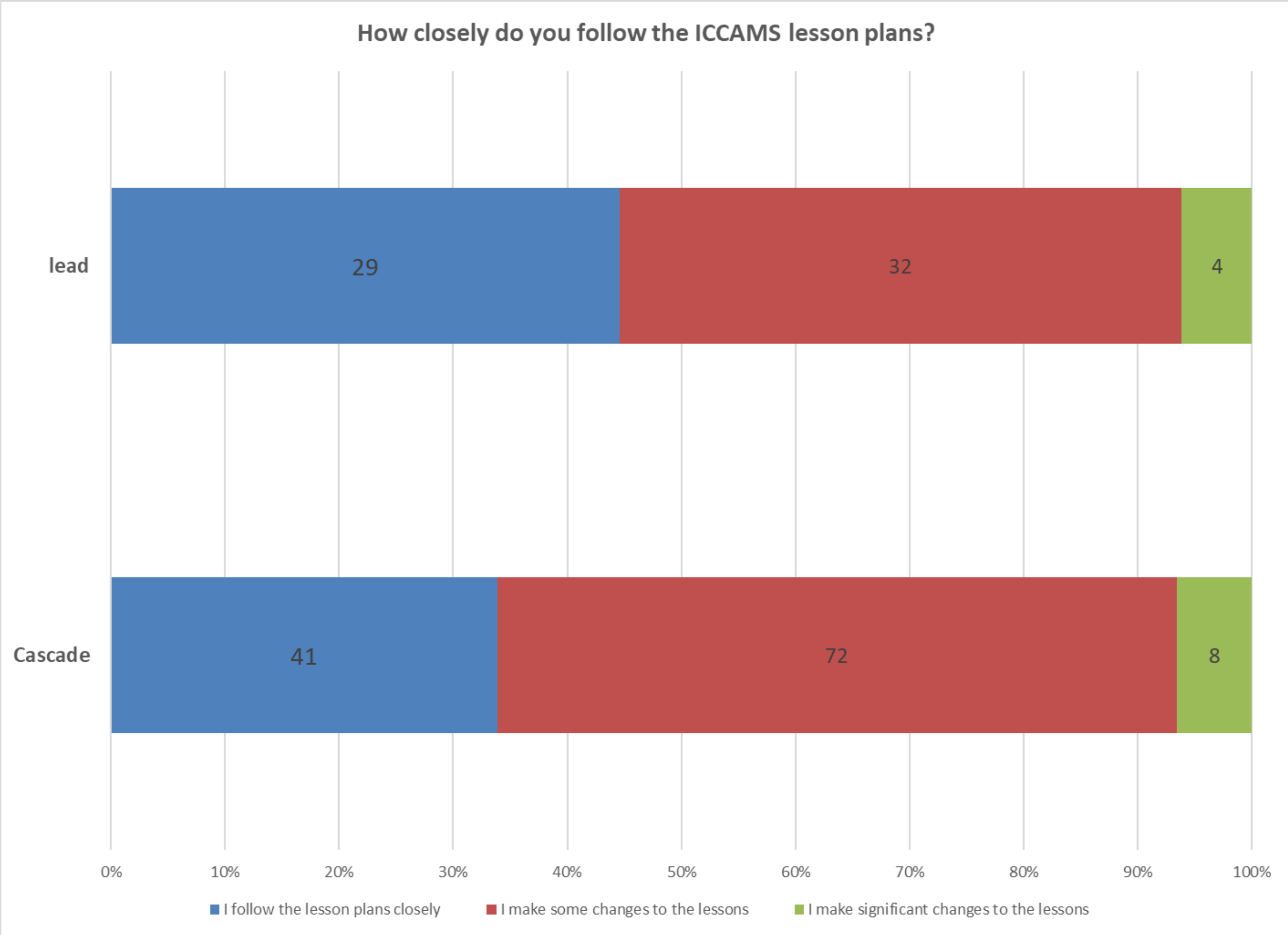


Figure 17. Distribution of teachers' responses to the question "How closely do you follow the ICCAMS lesson plans"

The above evidence reveals some concerns about the material and how the formative assessment approach is presented and also perceived by the teachers through the handbook/ICCAMS material. Interestingly, guidance in answer to the FAQs further suggests that the teachers might only need to read the first page of the notes (a lesson outline) though they are also encouraged to read the "overview" page also, in planning for the first teaching of the lesson (see also Figure 16 and quotes from teachers at the end of this section). The handbook states the "other materials may be more useful when you are teaching (the) lesson for the second or third time" (page 6). These other materials more directly address anticipations of pupils' responses to the tasks, and the developmental or learning trajectories anticipated in classroom dialogue – the key part of reflective discussion that is key to learning outcomes from formative assessment. Thus, the handbook anticipates a learning curve for the teacher, as also noted in some of the reported difficulties by the teachers, which might require multiple experiences of teaching the lesson, leading to a deep understanding of the lesson and its formative potential. This may be a realistic perspective for the long term developmental path for some or even most teachers, but might be a problem for expectations of significant effects of formative assessment pedagogy in the short(er) term evaluation duration.

Then there is the connectedness of the lesson materials and plans with the mini assessments. The mini-assessments are intended to inform the way a lesson will be conducted, but it is not made explicit in the teacher's handbook how this connection should be used pedagogically. This is evidenced with the variations in the timing and length of mini assessments during ICCAMS lessons, as reported by the survey teachers and shown in Figures 18 and 19. Figure 18 for example shows a difference in the distribution of lead and cascade teachers responses in regards to the timing of mini assessments, with more lead teachers being closer to instruction which they had also experienced with the PD sessions, as opposed to cascade teachers with more frequent responses about delivering mini-assessments further from the actual lesson and even not delivering at all.

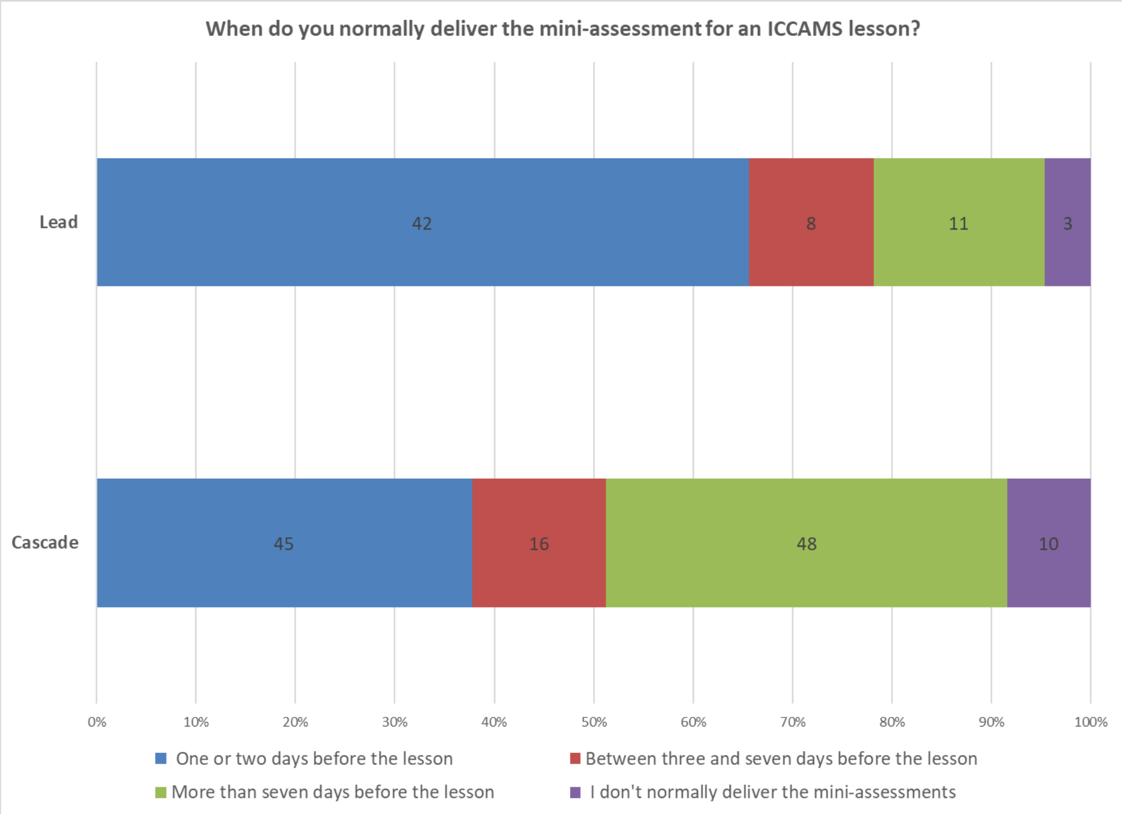


Figure 18. Distribution of teachers' responses to the question "When do you normally deliver the mini-assessment for an ICCAMS lesson?"

Figure 19 shows some further variation of the length of mini-assessment which could be a consequence of the direction about this aspect in the handbook.

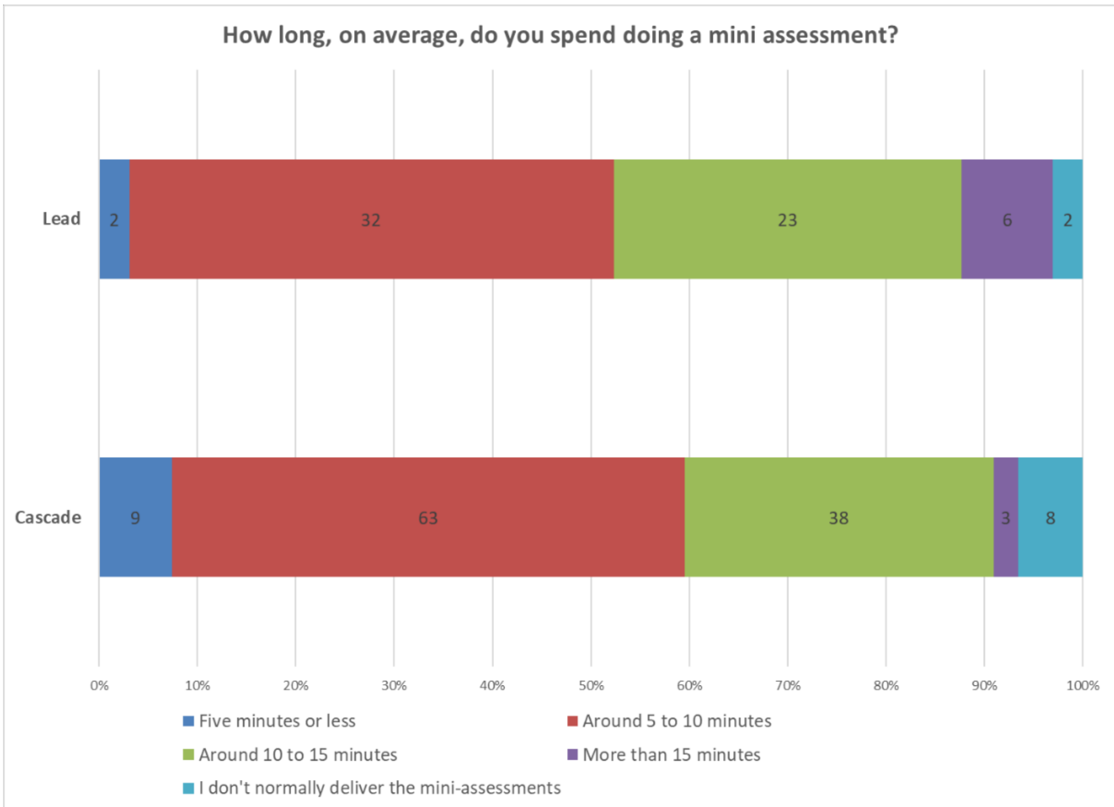


Figure 19. Distribution of teachers' responses to the question "How long, on average, do you spend doing a mini assessment"

This open-ness of the handbook about the pedagogical connections of mini-assessments with the lessons may have left some teachers confused, especially cascade teachers, as noted below:

> *" So I think, I don't fully understand what it [ICCAMS] is, and the research behind it. So I think, there should be more information. […] So… […]Yeah it tells you the outline of the lesson and then I'm trying to gauge it based on that, and I'm trying to pick up on what misconceptions it is trying to highlight and things."*
>
> (School 107, Cascade teacher, male, TY7=Medium, TY8= Medium, FA7= Medium, FA8= Medium, IC=Low, ID: S107T011)
>
> *"my thing with all the lessons, I get where it starts, the start of the lesson, one of the problems is then the structure of how you develop the lesson further. You know I read through this and sort of the power points, I feel are slightly left open ended."*
>
> (School 84, Cascade teacher, female, TY7=High, FA7=Low, ID: S084T001)

This was also noted during our observations in PD sessions: we observed participants asking the PD what to do with the pre-test results in one of the first sessions.

Further guidance and structure of the lessons in general was mentioned by other teachers as shown with the examples below, the former from a lead teacher who proposes breaking down the material.

> *"One problem we did have with it, because the resources were difficult to work with…[…] So definitely just breaking it down more for them. […] Yeah we've done a lot of work, so for one ICCAMS slide we turn it into 10 slides. So really breaking it down. Yeah, it would be massively helpful."*
>
> (School 41, Lead teacher, female, TY7= Medium, TY8= Medium, FA7=Medium, FA8= High, IC=Medium, ID= S041T008)

A cascade teacher from School 41 further commented on how they would welcome some concrete examples of how conceptual understanding can be assessed after some lessons:

> *"But I also think the materials could maybe guide how you would like us to scaffold that. Because I worry that I've scaffolded it so much sometimes, that what was the point?….[…]*
>
> *And then maybe even a way to see, 'this is an example of, in these 4 lessons, one of the things that would be amazing in the progress of this. Just so teachers have more of an idea of what success is in ICCAMS. What does look good for these students and what are we looking for them to build, other than conceptual understanding. Obviously conceptual understanding…. how can we see that, and how can we assess that. So I think having that in some way or another. Would be hard to plan but good.*
>
> (School 41, Cascade teacher, female)

Implementation of ICCAMS in schools

In this section we explore the evidence of implementation of ICCAMS in schools under various themes in relation to classroom practices and teachers' and students' experiences and perceptions. Evidence herein is provided in response to RQ6(b), RQ7, and RQ8.

School level—leadership perspective

From an engagement perspective, the intervention worked well when the two lead teachers in a school attended the regional sessions and were proactive in maintaining the intervention activity back in their own classrooms and in cascade PD. The intervention needed them to attend regional PD and bring its messages, through cascade training, back to the school; at best, they were acting as agents of the developers and PD leads through their constant presence in the school and not just in formal cascade meetings. One interviewed PD lead (Region 2) said it helped if senior management was supportive of the project and showed understanding of the needs of the mathematics department in their school priorities. It was the responsibility of the lead teachers to argue for, and find time for, the cascade training. Cascading the material as requested by the developers and trainers was one area where many schools fell short, failing to meet the project intentions (please see Fidelity for more detail)—for example, 20 minutes for cascade training was the norm, as opposed to the hour requested (which, as suggested above, provided a weakness in comparison to the time allotted for the regional sessions lead teachers attended).

Departments that most struggled to implement ICCAMS were often those that said they had either wider school issues that demanded teacher time and priorities, problems managing student behaviour, or struggles with staff retention, unfilled vacancies, and recruiting good mathematics teachers. This was voiced by PD leads and others from the delivery team. For example, the PD lead of Region 5 mentioned in particular:

> *'Where it was less effective was where senior management or the subject leader was not as committed or supportive. Top down isn't sufficient to secure real change unless you've got support at all levels. Sometimes it's senior leaders: 'No, sorry, we need everybody in the classroom' or, 'I'm sorry the budget just can't stand it' (PD lead, Region 5).*

The PD lead trainer from Region 2 also raised a similar issue and said finding good mathematics teachers was an issue for some schools.

Teacher attitudes, experience, and self-reported practices

In terms of effective classroom teaching, the conditions for success in lessons we observed included asking the right questions at the right time, managing the classroom, and understanding the essential mathematics and its formative assessment potential in the lesson.

One of the PD leads outlined their view on the processes used by effective ICCAMS teachers:

> *'[They should be] comfortable with an opening of the discussion, asking a challenging question and being able to accept a range of answers, being able to think very quickly, what's the next question… they can record it … and then move on to the next question' (PD lead, Region 5).*

Overall, the teacher surveys with intervention teachers revealed the following correlations between teachers' confidence in teaching with ICCAMS and their teaching practice (that is, 'transmissionism' and 'FA' practice) measures at Year 8: a negative correlation was observed with transmissionist teaching and a smaller, positive correlation with FA practice (Table 35), both very unlikely to be null. All these measures are teachers' self-reports, which is suggestive of some teachers' perceptions or understandings of connections between transmission, formative assessment, and the ICCAMS approach. It should be emphasised that these relationships are associations and we do not claim causal effects: less (self-reported) transmissionist teaching and more FA practice are associated with more self-reported confidence of teachers teaching with ICCAMS. This could be, for example, either that less transmissionism or more FA practice causes more confidence, or that more confidence leads to less transmissionism or more FA practice, or that there are feedback loops involved. Such possibilities are discussed later.

Table 35: Pearson correlations between teacher scores on confidence with ICCAMS and teaching practice measures at Year 8 (n = 188)

	R	Sig (p-value)
Transmissionist teaching (Year 8)	-0.289	0.0001
FA practice (Year 8)	0.180	0.0134

Further correlations from teacher surveys with the measures of ICCAMS confidence and other teacher practices with students' outcome measures (average) did not reveal any significant relationship. (We will revisit this later.)

The remainder of this section explores various aspects of teachers' experience with ICCAMS including their understanding of ICCAMS aims and emotional perspectives and experiences.

(a) Teacher understanding of ICCAMS aims

Lead teachers were more likely than others to describe the ICCAMS approach in terms that reflected the intention of the developers, referencing how ICCAMS helped teachers to identify misconceptions and make connections between areas of mathematics. These teachers were more aware of the aim of group interaction, and the way verbal feedback of ideas from students to the class can be used to resolve difficulties and misconceptions, whilst introducing ideas and mediating the discussion. Widely understood by most teachers were the roles of collaborative learning and discussion between pupils, enabling students to talk through their own ideas, along with use of multiple representations.

Well understood aspects of the intervention design also included the aim to help with problem solving issues, specifically in ratio, proportion, structures of multiplication, and the use of real-world applications. Teachers were aware that reasoning and communication were both areas where students struggle. ICCAMS was presented as being about conceptual understanding of mathematical ideas, and encouraging students to look beyond just finding the answer to solve the problem, pointing to some of the crucial metacognitive aspects of a successful FA intervention as indicated in previous literature. But it is hard to evaluate the extent to which the whole cohort of intervention teachers understood this and practised it.

A teacher in School 82 mentioned how ICCAMS allowed students to compare different methods and spoke of how this benefited the whole classroom dialogue. They often described classroom practices—peer learning, managing discussions, and questioning effectively—that were central to the ICCAMS approach. We are less sure of the learners' depth of engagement and reflection on these dialogues, thought to be crucial to the internalisation of the conflicting methods, conceptions, and learning strategies under discussion.

Two teachers, from School 4 and School 107, described the main principle as being a focus on one concept at a time, and that a whole lesson was built on that concept, with all the different mathematical ways and methods of understanding the problem being considered. Another teacher spoke of this focus on concepts:

> *'[ICCAMS] encourages understanding that should carry over time. Because it is developing more than just remembering a process. It's actually conceptual understanding' (School 41, Lead teacher, Female).*

We believe most lead teachers recognised that the lessons were meant to give students the opportunity to think and work collaboratively whilst also being structured and directed towards learning outcomes.

Lots of teachers said they found managing such a discussion hard, partly because they had to get used to sifting and taking forward productive ideas given by students ('Weeding and sowing', Freudenthal, 1978). Many, especially lead teachers, claimed they gained confidence with experience, as witness the survey results presented earlier. The discussion element was acknowledged as important, and with it an ability or disposition to see the value of questions and move away from a focus only on answers. One teacher from School 82 encapsulated this idea as follows:

> *'It's okay to put that bit on hold, extend this bit here. Because if someone asks a question, we need to know why' (School 82, Cascade teacher, male TY7= Medium, TY8=High, FA7=Low, FA8=Medium, IC=Medium, ID: S082010).*

The deeper level of understanding demonstrated in the classroom was mentioned by a small number of teachers during interviews. One spoke of cognitive conflict, although not in precisely those terms, when describing how students were challenged in each lesson with something they did not understand and how discussion of conflicting ideas was used to help student understanding. Some teachers spoke of the value of reading and expression through discussion and writing (which were particularly valuable in revealing what students were thinking):

> *'ICCAMS … it's a lot more focused on the reasoning, and that deeper mathematical understanding … the discussion bit, the oracy, I think is a really important stepping stone on that journey' (School 82, Lead teacher, male, TY7= Medium, TY8=Medium, FA7=High, FA8=Medium, IC=Medium ID: S082T004).*

Representations were sometimes noted as an important aspect of ICCAMS. One teacher said that that they 'enjoyed teaching pictorially' and that the use of diagrams and models, and the encouragement of pupils to draw, was helpful in engaging students and furthering understanding. The ratio tables and double number lines were mentioned as particularly useful tools for problem solving.

This raises some questions about teachers' pedagogic understandings, that is, about how concepts like 'understanding', 'dialogue', 'models and representations', and 'problem solving' connect with formative assessment: in fact these terms, especially terms like 'metacognition' and 'cognitive conflict' were rarely if ever used in the interviews and discussions we observed, perhaps because these critical concepts were not explicitly addressed in the ICCAMS design and implementation.

Some teachers, during interviews, compared ICCAMS to their work with the 'Mastery' programme or curriculum. A teacher from School 4 said the ICCAMS-style lessons fitted well with Mastery, with the shared features of in-depth teaching and subject knowledge (see more on this below). Another teacher from the same school highlighted the differences between the two approaches, noting that Mastery was perceived to be more structured, and more focused on breaking down skills, whereas ICCAMS perhaps has a stronger conceptual focus. As will also be detailed with regard to the case study, this may suggest that some teachers are reflecting on the contrasts between the principles and theories involved in different developments and interventions and thus attending to the theory or praxis and not simply compliance or delivery of policy.

(b) Positive experiences of teachers

Many teachers interviewed said that the intervention had made them think differently about how they teach. They reported that it had made them different teachers, more comfortable with dialogue in the classroom and using contributions from students to help classroom discussions and hence develop understanding.

Teachers were more likely to be positive when they felt the ICCAMS approach aligned with their teaching style.

> *'I've really enjoyed being able to teach more in my style of facilitation rather than teaching to the scheme of work/ preparation for assessment' (School 15, Cascade teacher, female, TY8=Medium, FA8=Medium, IC=High, ID: S015T013).*

But there was also evidence of new teaching practices leading to new understanding amongst a broader range of teachers. Some teachers said their teaching had changed, not just in Years 7 and 8 but across different year groups, indicating a deeper change in pedagogy and not one confined to the ICCAMS lessons given to be 'delivered'. The changes in pedagogy were often explained in terms of the nature of student initiative and student-teacher dialogue:

> *'I take more of a back seat now. As much as possible I try and get the kids to do the thinking and the explaining' (School 98, TY7= Medium, TY8= Medium, FA7= Medium, FA8=Low, IC=High, ID: S098T006).*

Lead teachers were more often likely to say their teaching had changed (see also Figure 12). Some felt the project had opened their mind to ways of working that they had not considered, or had forgotten after getting stuck in teaching routines.

The teacher survey data also provide evidence of regional differences in confidence scores and also that in all regions lead teachers consistently scored significantly higher than the other (cascade) teachers in their school who completed a survey at the end of Year 8 (see Figure 20). In almost all cases, the difference in favour of lead teachers is notable in effect as well as significance. Further differences across regions, but also between cascade and lead teachers within each region, are presented in Appendix 22 (Figures 22A and 22B).

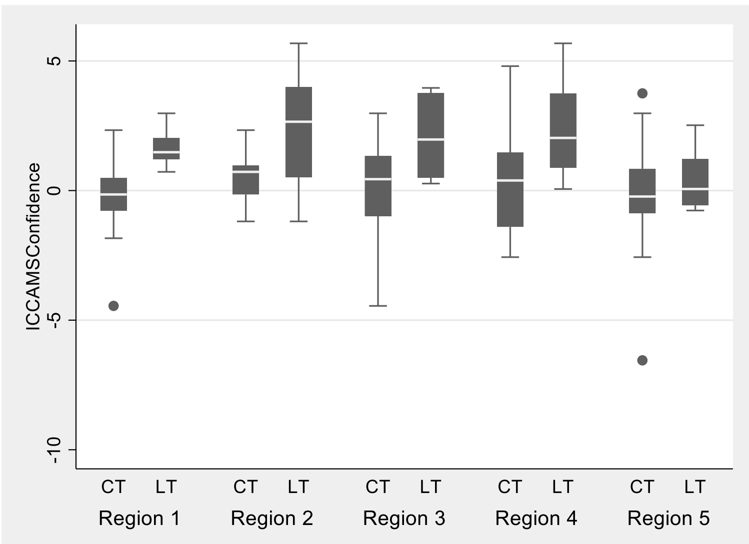
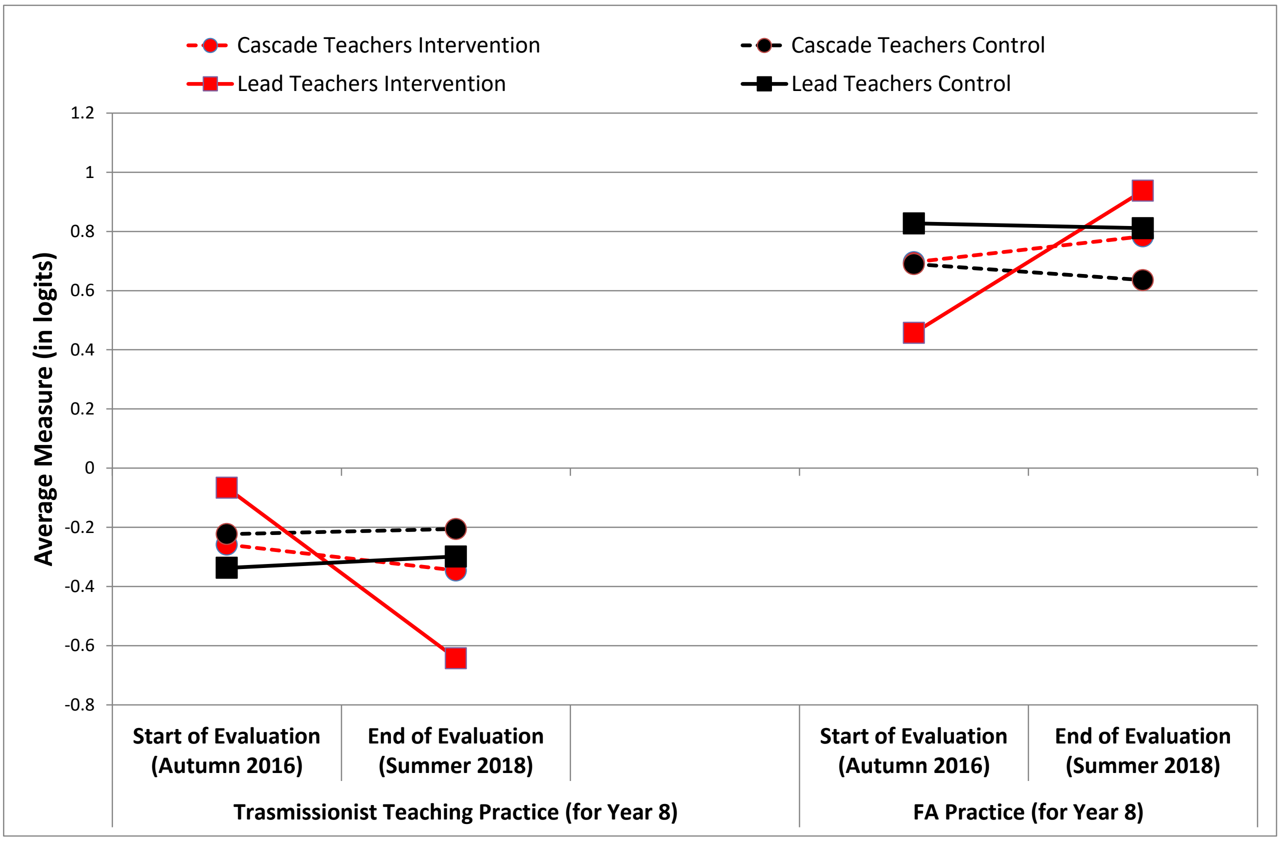


Figure 20: Boxplots of confidence with ICCAMS by teacher type—lead or cascade—and region

From the teachers' perspective—as already mentioned in relation to Figure 12 at the start of this section—we can see a decrease in transmissionism and increase in FA practice. Any such decrease between Year 7 and Year 8 is likely to be due to 'bedding in' of practice as the project progresses.



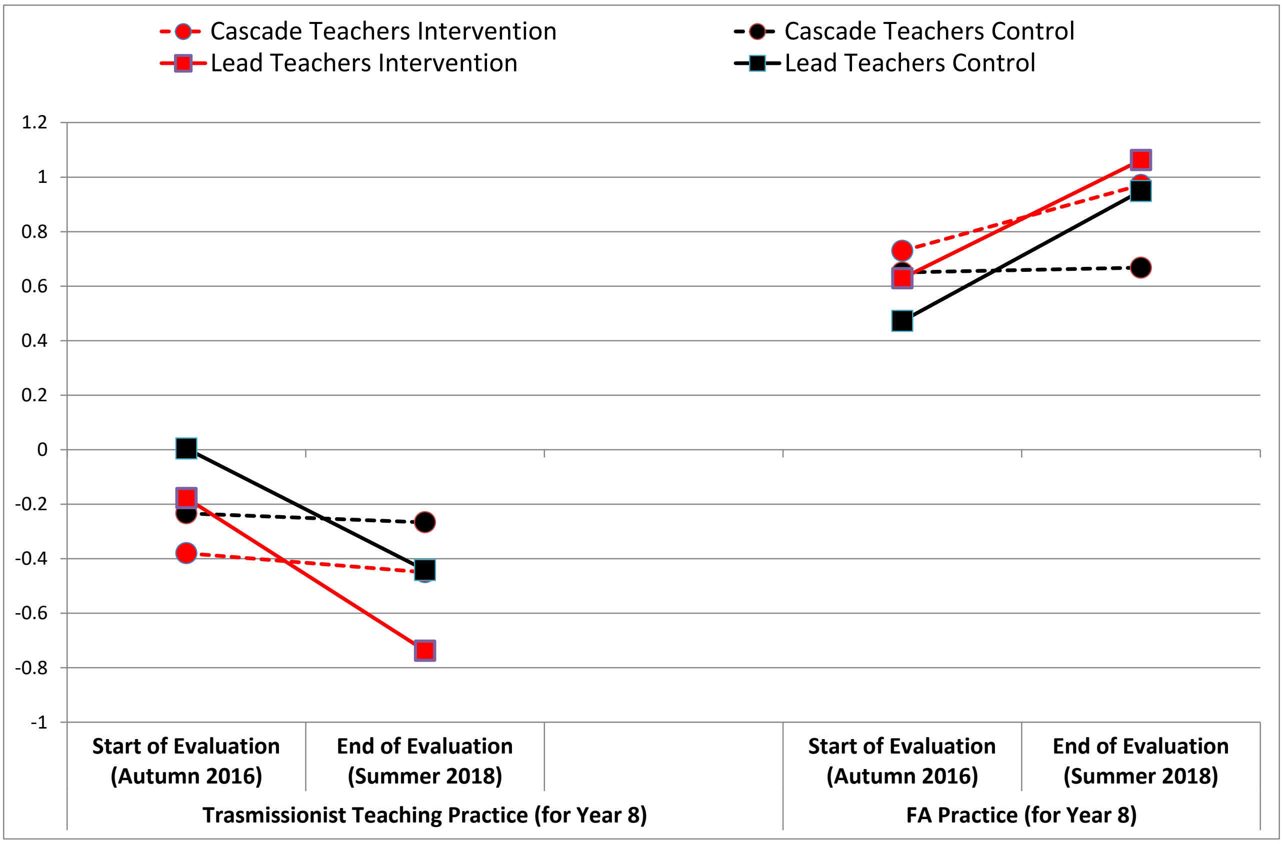


Figure 21: Changes in teachers' perceptions of transmissionism (left side) and FA practice (right side) by teacher type and condition—cascade/lead x intervention/control (top: all available responses; bottom: only matched responses from 179 teachers—intervention: 44 CT and 50 LT; control: 59 CT and 26 LT)

From Figure 21, it seems only the lead teachers changed significantly over the period of the intervention in terms of their self-reported formative assessment practice (increasing) and transmissionism (declining). Similar effects were noted for cascade teachers, especially when considering the matched sample (bottom figure) in control schools, which might be a reflection of control school lead teachers also working on their pedagogy (for example, their engagement with programmes such as Mastery as noted below). An alternative interpretation, especially considering the cascade teachers, could be that differences may be due to differences in the samples from DP1 to DP2.

There is also a strong indication of change in perceptions of teaching practices based on student reports as shown with the results in Table 30 presented earlier.

(c) Perceived challenges reported by teachers

The frequently perceived teacher difficulties (mentioned in interviews) when teaching the ICCAMS lessons were:

- students disengaged and classroom management issues;
- low ability students found it tough, requesting help or needing 'scaffolding';
- teachers confused if pupils answered quickly, and could not move lessons forward;
- reluctant staff, some found the 'open' nature of tasks stressful; and
- departmental issues—staff disengagement and staff turnover.

Teachers noted perceived weaknesses in the intervention and described several struggles with it. We found the ICCAMS approach was sometimes dismissed when a **teacher did not feel aligned with its principles** suggesting the success of the intervention depends in large part on the prior beliefs and practices of the teachers. In one school, a lead teacher told us of one of their (cascade) colleagues who had conducted an ICCAMS lesson in silence. Reasons for these struggles include a dislike of the lesson structures and some teachers preferring more transmissionist 'delivery' approaches with less student collaboration and discussion. Some experienced teachers were said to be 'set in their ways' and more negative about changing their practice, particularly as the intervention seemed to demand a new style of teaching about which they lacked confidence. However, many of the most experienced teachers were also willing and able to learn new things from the approach, notably in regard to generating discussion and thinking about problems in different ways.

Some cascade teachers appeared not to **understand what the intentions of the intervention** were and seemed to miss the point of some lessons we observed, for example, saying there was not enough content to fill a lesson and that lessons were unstructured and too open-ended. Interpreting this, we concluded that these teachers had not fully understood, but also did not realise what changes in practice were intended, or **did not get sufficient support** to implement the intended new practices.

According to the PD lead of Region 4:

> *'Some of the cascade teachers, particularly the first time round, left it so loose, they were very aware this was a different type of pedagogy to what they were used to … They were trying to make everything a discussion point. And so the lesson lost momentum. […]*
>
> *There was quite a lot of, "I know in an ICCAMS lesson, I mustn't give you the answer." [This] was one of the myths that came out and so they were trying to draw everything from the kids' (PD lead, Region 4).*

There were many comments about the **difficulty of making ICCAMS work for the lower attaining classes**; we interpreted this as a need for 'scaffolding', especially for 'weaker' pupils. This was a typical comment highlighting such sentiment, expressed as an open response in the survey:

> *'Difficulties were the restriction of the direction of the lesson and the style of delivery. Barriers [included] the level of mathematical content for weak students' (School 87, male teacher, cascade, TY8= Medium, FA8= Medium, IC= Medium, ID: S087T009).*

For some, the open teaching style, perceived by teachers as the ICCAMS-favoured style**,** was often claimed to be **difficult for already disengaged students** and then **behaviour management** became a problem. Our lesson observations confirmed that teachers sometimes struggled to get the students to share their ideas with the group, struggled to limit their own contribution, and in such cases the discussion came to a conclusion too quickly. Such difficulties are commonplace in PD projects developing more discussion-based or dialogic pedagogies and has been remarked upon many times, most obviously in comments on the demand of the plenary of the ‘three part lesson’ but also internationally—even in Japanese lesson study where experts say that the plenary ‘neriage’ makes a high demand on teachers’ skills (Fernandez and Yoshida, 2004).

But then also, a lead teacher in School 82 said that although teachers in the department with a more traditional style had found the project challenging, it had helped to bring a culture of discussion to classrooms. Two lead teachers we interviewed in this school said that teaching ICCAMS becomes **easier with time**. They felt students got more out of it as time went on and as their teaching adjusted. It should be noted that the ‘easier with time’ comment was mentioned by both lead and cascade teachers. This department, as well as the one from School 4, had started teaching some ICCAMS lessons to other KS3 and KS4 classes: this ‘spreading’ of pedagogic practice was mentioned in several schools. Both these schools were in Region 5 and had a relatively high fidelity score of 1.08 (approximately the upper quartile, well above the average of 0.11 and at the edge of the medium/high group; see Figure 13).

Many teachers, as for example in School 84, reported that at the start of the project they believed the ICCAMS lessons in the handbook would not fill an hour yet teachers often found they were soon able to make the exercises fill a lesson at least. Team teaching occasionally took place in this school when teachers had had reservations about their ability to teach the lessons (often they said these reservations were because they had a teaching routine that they trusted and did not want to change, commonly expressed as resistance to any reform of practice, as it relates to the teachers’ sense of efficacy). The lead teacher there said in interview that she joined cascade teachers in their classes to help overcome their doubts.

One teacher explained that they liked the focus on thinking processes, as opposed to ‘yes or no’ solutions, but said the quality of discussions depended upon the class and their willingness or disposition to work collaboratively. This teacher felt the lessons were left slightly too ‘open-ended’, without closing on the outcome they were ‘hoping for’:

> *‘Just depends on what the kids come up with. And sometimes if they don’t quite see the links that you are hoping that they do …’ (School 84, Cascade teacher, female, TY7=High, FA7=Low, ID: S084T001).*

Some teachers admitted finding the lessons **harder to teach than normal mathematics lessons because there was less control**. This was evident in School 107. Some staff were more reluctant than others, particularly if they had a fixed idea of their own dominant teaching style. One young teacher responded as follows when asked about difficulties in relation to difficulty managing discussions:

> *‘[It’s difficult] trying to follow the students’ logic whilst you are kind of trying to manage the classroom, I guess’ (School 107, Cascade teacher, male, TY7=Medium, TY8= Medium, FA7= Medium, FA8= Medium, IC=Low, ID: S107T011).*

This particular teacher seemed not to understand the intentions of the project and had received little cascade training. The teacher said they had looked at the outline of the lesson before the class and just went from there. They said they were sometimes unsure of the misconception they were attempting to address.

One lead teacher in School 107 told us they enjoyed teaching ICCAMS and said it was different from their usual teaching, highlighting the use of different methods to solve problems and find answers. They also mentioned it became easier over time, after teaching a couple of lessons in depth and getting a feel for them (see above on this point). But they also said ICCAMS could be ‘exhausting’ because of the need to also focus on behavioural management:

> *‘You have to be constantly making sure that they are actually doing something productive … making sure that chat is useful, and not just nonsense’ (School 107, Lead teacher, female, TY7=Low, TY8=Low, FA7= Medium, FA8= Medium, IC=High, ID: S107T009).*

Some said their **department was not so well suited to the intervention**. Tensions were sometimes due to a different pedagogic tradition. School 41 was one example. The teachers here also mentioned behaviour management was particularly important in ICCAMS lessons. Their teaching appeared traditional, with transmissionist teaching methods prevalent, and they lacked confidence in, and fidelity to, ICCAMS (school fidelity: -0.99; average teaching transmissionism DP2: 0.27; average ICCAMS confidence: -1.79, based on teacher surveys). One teacher there said that their lower ability students could not access ICCAMS without ‘scaffolding’ and hints and said ICCAMS had not increased student confidence, because ‘direct instruction’ was being used in the ICCAMS lessons. Another colleague admitted difficulty in facilitating discussion-based mathematical activities and suggested students would have perceived their teacher's lack of enthusiasm for ICCAMS. This suggested the worrying possibility that teachers who had little confidence in the intervention, while still engaging with ICCAMS, might produce unintended ill-effects and poor outcomes. This in turn raises questions about school commitment and decision making in such situations, which is revisited in the discussion.

One teacher in School 41 expressed some ambivalence: they said that students sometimes need to work independently, stating that:

> *'ICCAMS [is] sometimes … a bit too much … you wouldn't want that every lesson.'*

However, the same teacher also said:

> *'It did really make me realise, actually, that giving more time to explanations … is so beneficial for so many of them … they do need to check their understanding in different ways' (School 41, Lead teacher, female, TY7= Medium, TY8= Medium, FA7=Medium, FA8= High, IC=Medium, ID= S041T008).*

This could be indicative of teachers who were in the process of development. We consider that there might be (1) those that saw the development as already aligned with what they believed or practised, (2) those who were more or less strongly misaligned (and likely to be disaffected), and then (3) those in between who were open to consider ICCAMS and associated changes but who might be sceptical about aspects of the changes being sought.

(d) The context of 'ability' and low attainers

A frequent statement by teachers was that the intervention lessons were too difficult for low-attaining students. Teachers often felt the need to adapt the materials in such cases in particular. Some said such pupils could not engage with ICCAMS because of their difficulty with simple calculations. In School 81, one teacher attributed the problem to disengagement and a lack of curiosity with the subject. However, another teacher in the school said that some lower attaining classes responded 'very well' to the ICCAMS approach:

> *'I think it's worked very well with the lower ability groups in this school, where it's a good way of exploring maths ideas in a different, non-traditional way' (School 81, Cascade teacher, male,TY7=High, TY8= Medium, FA7=Medium, FA8= Medium, IC= Medium, ID: S081T009).*

A sentiment expressed during the teacher interviews by some teachers was that the lessons were too hard for weak classes and too easy for higher tier classes. However, it was also said that some higher attaining students were less likely to want to ask 'why?'. One teacher mentioned that higher ability students usually answered very quickly and, at first, this teacher was unsure how to then further develop the lesson. Yet this teacher also said that with time, both teachers and students started to appreciate the work and how it could 'stop them in their tracks' at times:

> *'It really stops them in their tracks because they were used to just being able to do something … But as time went on, I think they really, really started to appreciate it' (School 82, Cascade teacher, male, TY7=Medium, TY8=High, FA7=Low, FA8=Medium, IC=Medium, ID: S082T010).*

At the PD days, there was discussion of how the lessons could be tailored to students of differing attainment. It was acknowledged in discussions that some classes would not progress as fast or as far as others and may need additional resources, and it was understood that teachers had to meet the needs of the students whilst adhering to the principles behind the project.

This kind of discussion of conflicts for pedagogy seems to us to be essential as part of the PD process for teachers (just as it is for students in classroom mathematics), but we are not confident that this often occurred in school cascade PD meetings. On this point, however, as with much else, the absence of evidence should not be taken to be evidence of absence.

(e) Formative assessment

One of the key ICCAMS intentions was to help teachers practise formative assessment. In School 84, a lead teacher said the mini-assessments helped, although it had been a struggle to fit them in because of time. A cascade teacher admitted they were not good at doing the mini-assessments because pupils could not always start them without initial direction. They had not done all the mini-assessments for this reason. One might conclude that for this school, even though it topped the fidelity measure, there was a failure to grasp the essentials of the project (an important concern with the external validation of the fidelity measure adopted in the quantitative analysis of fidelity effects).

A similar feeling was expressed often in School 81—the lowest scoring case study school. Limited time and student struggles were said to be problems, but one teacher mentioned they used the mini-assessments and when they generated interesting responses they were used to ‘tailor lessons’, perhaps expressing an essential principle of formative assessment without naming it as such.

> *‘I treat it as a bit of “that’s what we got to do” … if I do receive anything interesting, then I will use it to tailor the ICCAMS lesson. But often it feels hard … Often, they don’t perform so well in it … I realise I just need to teach that lesson’ (School 81, Cascade teacher, male, TY7=High, TY8= Medium, FA7=Medium, FA8= Medium, IC= Medium, ID: S081T009).*

Teachers in School 82 (a medium fidelity school) also used the formative, mini assessment exercises to determine whether to amend the lessons. The lead teacher here expressed surprise that the mini-assessments were intended to be conducted completely separately from the lesson and said that pupils did not like it when they did not go through the answers. This raises the question of learners also needing time to understand the ICCAMS process, get used to, and value the new lesson regime (even when it is not consistently practised). Added to the time-lag for teachers to feel they were becoming proficient in teaching these lessons, this might also help to explain why sometimes a short term dip in performance might precede longer term benefits.

In School 41, a teacher that admitted struggling with ICCAMS said that they did not know how to see and assess what students should be achieving. When asked about the formative assessment, they said:

> *‘It’s quite hard to see what success would look like. It seems very much, “this student understands this problem”, or “this student did not understand this problem”’ (School 41, Cascade teacher, female).*

From our interviews with teachers and survey responses, it seems the formative assessment principles were not as widely understood and engaged with as the associated but broader principles of discussion and collaborative working (this latter being but a part of the requirements of the former). This is not to downplay the importance of what the intervention aspired to, or even perhaps achieved in such a short space of time. We will return to this later.

To conclude: the pedagogy of using task specific feedback with wider cognitive ambitions in formative assessment was perhaps not achieved to the extent it might have been. We suggest this was perhaps because this aspect of formative assessment lay outside of the main *principles and terminology* presented in the handbook and taught in the PD, and manifest in lesson objectives and plans. Its use would have required a deeper understanding of pedagogic content knowledge, ‘understanding in practice’ and metacognitive principles rather than being reduced to ‘discussion’ alone (see Shulman, 1986).

Student attitudes and experiences

This subsection provides evidence in response to RQ8. In positive cases we noted there were signs of increasing students’ confidence in tackling new, more open problems, that is, when ICCAMS allowed students to share and discuss ideas and encouraged them to understand mathematics conceptually rather than being reduced to remembering procedures.

Benefits to pupils mentioned by teachers in interviews and at the observed PD sessions included better reasoning and improved competence and capability to express ideas in context. Teachers frequently mentioned that ICCAMS had led to more pupils thinking for themselves and expressing themselves verbally. On the whole, teachers tended to say the intervention improved student’s oracy, or ‘speaking and thinking’.

Indeed, one teacher said ‘many students were unwilling to have a go—this improved the more ICCAMS lessons they took part in’. This statement was representative of a widespread feeling amongst teachers’ celebrating ICCAMS: it was all about learners having a go and improving students’ confidence and feeling that they could do so without negative effects. This is the counterpart of those teachers noted above who saw pupils activity as a behaviour problem undermining classroom management.

In School 4, the teachers said that ICCAMS had reinforced independent thinking, with students helping lead the lesson with their contributions. They said the project had fostered a ‘can do attitude’ amongst students with them more confident trying different problems, offering their thoughts, and attempting new methods. This expression and lack of fear of being wrong were things these teachers were now trying to build with other year groups. In addition, teachers said they have learnt to pause and listen, rather than jumping in straight away, in order to give students time to think. (See more on this school case below.)

This sentiment was shared by a teacher in School 98, saying that they try to minimise their role by—

> *‘not jumping to the conclusions for them, letting them come up with all the reasoning and trying to link the reasoning together to get to their conclusions. And I’m almost now more … I am happy if there is no conclusion that they get to’ (School 98, TY7=Medium, TY8= Medium, FA7= Medium, FA8=Low, IC=High, ID: S098T006).*

One teacher also said when they had a good appreciation of the lessons, they felt they could lead with student ideas and guide the discussion to a place that helped the class. But another teacher said that at the beginning, the lessons were quite abstract for students and they needed to overcome the feeling that they were incompetent because they did not understand the material.

In School 84, one teacher reported the same ‘can do attitude’ to formative assessment and self-regulation. In a model of what ICCAMS might have hoped for, they said:

> *‘They are not just saying, “I’m stuck, I don’t understand?” They are specifically telling me what they don’t get. And that’s a big eye-opener—they are exactly saying what they are struggling at, and I’ve never known that before. If they can identify their weaknesses, then they know what to work on to improve’ (School 84, Lead teacher, female,, TY7= Medium, TY8=Low, FA7= Medium, FA8= Medium, IC=High, ID: S084T007).*

Teachers in School 41 said that students were not used to working in groups but that collaboratively discussing the mathematics had helped to develop their understanding. One teacher there said they were surprised how much ICCAMS had engaged certain students who did not do well under traditional methods and assessment:

> *‘Some surprised me that they were better at the ICCAMS than they are in the traditional assessments and things. [They] get to grips with certain ideas from the ICCAMS … but maybe when we do assessments here they come down the bottom of the class. So for them it was a nice confidence boost to think, “oh I can actually …”, kids who can think outside the box, when actually kids are, like, “give me a method … I just want to follow a method”. I’ve got such a mix of either: really love it, or not a big fan’ (School 41, lead teacher, female, TY7= Medium, TY8= Medium, FA7= Medium, FA8=High, IC= Medium, ID: S041T008).*

According to some teachers, students saw ICCAMS as ‘too much of a fun lesson’ or, in the words of another, a ‘doss lesson’ requiring fewer written exercises. A lead teacher in School 107 said it is hard to know what students were doing when in group discussion (as noted above, they could just be chatting). Another teacher in the school was concerned that when another student talks, others may not understand and switch off. Some teachers also said they needed advice on how to ensure students have meaningful discussions. Although the group work and interactive element was liked by most students, as mentioned above, some teachers complained about this generating behaviour management problems.

In interviewing students and engaging with students in class, some described the lessons as ‘more interactive’ and ‘more relaxed’. They were aware the lessons were different from their normal lessons, usually explaining this is terms of their feelings and level of engagement. Notice here there was little evidence of the learners being explicitly introduced to a new didactical contract in terms of formative assessment. But one student spoke of deeper mathematical understanding and in regard to the ‘contract’ added:

> *‘You had to use your own initiative and work out what it was saying with [them] giving it straight to you’* (*School 82, Student).*

Many schools had integrated ICCAMS into their scheme of working for KS3, and sometimes KS4. This is evidence that the teachers thought the resources and teaching practices were helping their students.

Looking at the effect of implementation of ICCAMS and improvement in students’ engagement and attitudes to mathematics, we further draw on evidence from the teacher surveys complemented with their students’ learning outcomes (as averages). A model (in Appendix 21) of average mathematics dispositions at DP2 (of all the students assigned to each teacher) considering average mathematics disposition at DP1, teacher type (lead/cascade), years of teaching experience, and teachers' perceptions of FA practice and transmissionist teaching at DP2 is indicative of the positive effect of the intervention, albeit with low confidence of this effect not including zero ($p = 0.053$). This suggests that the mathematics dispositions of the students of the intervention schools may increase at a faster rate than those of control schools, all other factors being the same (that is, controlled for).

Fidelity

In this section we revisit in a more targeted way the issue of fidelity and provide evidence of the extent to which ICCAMS was taught as intended. It must be noted that most of the issues we have already presented relate to fidelity: so far we have looked more into various stakeholders' perspectives and how they experienced the intervention. In this section we turn to the intentions of the developers and evaluate how various elements were enacted and we then revisit the overall fidelity measure and its relationship with other school- and teacher-related variables (including at a regional level).

Intentions

To repeat, the following actions were asked of schools involved in the intervention:

- teach all lessons to all classes;
- lessons delivered fully and as described;
- conduct mini-assessments;
- two teachers per school to attend PD training; and
- lead teachers to deliver hour-long cascade training after each of the nine PD training sessions.

Each of these listed intentions were, on occasions, considered or perceived as adaptable by teachers. Some schools and some individual teachers often did not implement the programme as intended.

Intervention lessons generally were compliant with the ICCAMS approach, according to our observations, but this was not the case in all schools or, of course, in all classrooms (see also Table 36). Teachers told us that their approach to ICCAMS lessons varied according to the prior attainment of the class, their interpretation of the lesson demand, and their practices or teaching 'style' as well as their commitment to the intervention. Some teachers were 'delivering' ICCAMS without much focus on intentions and their pedagogy might be described as 'business as usual' while complying superficially (for example, using some of the teaching materials). This generally applied to cascade teachers (as evidenced by observed lessons or interviews) rather than the lead teachers who had participated in the regional PD. In some cases, they knew that ICCAMS involved classroom discussion but there was not always as much purposeful 'led discussion' as ICCAMS intended (that is, directed to the conceptual issues each lesson demands) and they did not always draw ideas or 'threads' together from pupils' contributions. Our interview data illustrated how some of these teachers were also more likely to adapt aspects of the lesson plans and seemed less committed to the mini-assessments, whose essential contribution to the pedagogy was often not understood.

We conclude from this that the impact from simply picking up the lesson materials and 'delivering' them in the context of the usual pedagogy of the traditional teacher is unlikely to make the difference in learning opportunity that ICCAMS intended, and that the quality of the PD and its effect on classroom practice is an essential requirement for these intentions to be effective. There are many caveats to this conclusion, however, which we discuss below as we reflect on each of the predefined components of the fidelity measure (for the PD attendance, please refer to earlier section).

Frequency of implementation in schools

Some teachers did not teach all the lessons, reporting that they did not have time or that the content was too difficult for their pupils. In fact from the 193 teacher survey reports of teaching certain classes, only 80 (41.4%) reported that they taught all 20 lessons. In teacher surveys, it was mentioned frequently that teachers ran out of time to teach the lessons in the final stages of the programme. The biggest problem reported was that lessons were 'left to cascade teachers to provide [and] not all did so consistently' (lead teacher in school 54, from teacher survey). We do not have complete and consistent data on this as there was missing survey data from teachers, but a large number of lessons in intervention schools were, even according to self-report survey responses, not taught—as the 41.4% figure suggests.

Figure 22 displays data from those teachers who reported in the survey (DP2) which classes they taught ICCAMS to and how many lessons they taught (193 class cases from 45 schools). It shows that reports most commonly (n = 80 classes) say they met the aim of teaching all the lessons but on the other hand the majority of classes (n = 110+) were taught fewer lessons.

Most teachers said they adapted lessons. Of 65 lead teachers, 55% said they had made changes to lessons; in contrast, 67% of cascade teachers said they had not (see also figure 17). Perhaps the difference is best explained through 'confidence' in use of ICCAMS in the teacher survey (as reported earlier). Lead teachers were generally more engaged with the principles of the project, had the higher quality PD sessions, taught more lessons, and, according to observation reports, more were taught in the intended manner. In addition, we noted from the interviews that lead teachers began the introduction of the programme of lessons before the cascade teachers, and this also offers a possible reason for them to have gained confidence in the ICCAMS approach. In general, however, we cannot discount other explanations, for example, that our more experienced teachers might, on the whole, have expressed more confidence in their capacity to adapt and manage change,

Eight of 63 lead teachers said some Year 8 classes were not receiving ICCAMS lessons at all. In certain cases, these schools judged that some of their classes would not benefit whereas others said ICCAMS classes were being taught by external supply teachers and, thus, were not receiving the lessons.

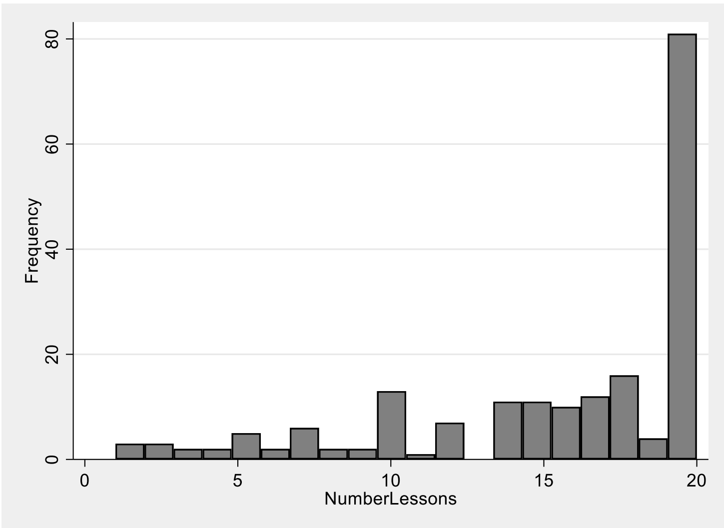


Figure 22: Number of lessons taught reported by each teacher self-reporting

Cascade teachers and cascade training

One problem observed in some lessons taught by cascade teachers was an over-emphasis on the role of 'free discussion', that is, discussing without achieving or intending to achieve a focal learning outcome and without there being a message to tease out and relay to the whole group of students. This is further evidenced by the two quotes provided by the PD lead of Region 4 presented earlier. Our interpretation of this is that there is a key difference between dialogue in the conversational sense and a mathematical dialogue with the collective intention of teacher and pupils better understanding the mathematics at issue (this difference is fully discussed in Ryan and Williams, 2007). It seems that the cascade teachers were particularly prone to this conflation in their understanding or practice of dialogue within the lifetime of the ICCAMS project.

Sometimes, as a consequence, discussion was used too loosely and the lessons lost focus. This could even be accompanied by the misunderstanding that in ICCAMS lessons the 'answer' to a problem should not be given to students (whereas of course this is a strategic decision depending on the context and assessment of students' understandings in the moment). However, as the project progressed and teachers got to grips with the aims, these misunderstandings were sometimes reported to be less frequent. One of the PD leads reflected on this during an interview:

> *'Interestingly on the ICCAMS project, one of the most effective parts of it in year one was when we went and watched the teachers teach because up until that point it was just talking to teachers and … sharing materials*

with them, trying to get them to be enthused by it, and they all responded, on the face of it, very well. We actually went and saw the teaching to give them some guidance … and feedback. And we did have time for more sensible feedback. And they got written feedback. That was really, really helpful. And I've noticed from the second year observations that teachers seem … they've definitely gotten better at it' (PD lead, Region 1).

This might be key: the effect of the PD may only develop over significant periods of time (time for ideas to sink in and thence to embed changes in practice) as was already reported in the review of relevant research evidence (see, for example, Adey et al., 2004; Gersten et al., 2014; Ruthven et al., 2017). We return to this discussion later below.

On the whole, lead teachers had a much better understanding of the intentions of ICCAMS in terms of style of pedagogy and mathematical aims (for example, over 74% on three items in Table 36, comparing well with their cascade teachers). Even so, some cascade teachers were very confident (see the 54% and 66% on those items in Table 36). We also observed some of this confidence in their relationship with students and their ability to communicate the material, which resulted in some strong ICCAMS lessons from cascade teachers.

Table 36: Lead vs cascade—alignment and confidence with ICCAMS

Statement	Lead teachers	Cascade Teachers
I feel confident about ICCAMS teaching	78% agree or strongly agree 2% disagree or strongly disagree	54% agree or strongly agree 20% disagree or strongly disagree
Feel need for further training	22% agree or strongly agree 59% disagree or strongly disagree	45% agree or strongly agree 27% disagree or strongly disagree
Teaching ICCAMS lessons matches my teaching skills	74% agree or strongly agree 7% disagree or strongly disagree	41% agree or strongly agree 19% disagree or strongly disagree
Would feel confident teaching again	85% agree or strongly agree 0% disagree or strongly disagree	66% agree or strongly agree 12% disagree or strongly disagree

The most frequent cascade training difficulties reported in interviews and survey comments by the intervention schools were:

- not enough time to meet;
- maths teachers busy with other responsibilities in the school;
- reluctance of other staff to engage;
- other curriculum demands;
- staff turnover/absence; and
- engaging new staff in the middle of the project.

The following sums up much of the reports from observations:

> *'The weakest link in the whole project, really, is how much time the lead teachers get to feed back down to the cascade teachers. And it's quite noticeable the difference in quality in the person that is taught by the lead teacher and sometimes the lessons are taught by teachers who kind of miss the point and [have] just taken the materials without the true understanding of the pedagogical processes that are going on' (PD lead, Region 1).*

The intention was that cascade training sessions would be an hour long. In the teacher survey, 35% said these sessions were less than half an hour, 41% said about 30 minutes, and 13% said an hour or more. Of 64 lead teachers, 45% said all cascade sessions had been taught and, of 125 cascade teachers, 67% said they had attended all training sessions offered. This suggests about 30% of cascade teachers had the intended full experience of nine sessions while less than 10% of their sessions would have been an hour long. This can be compared with the lead PD all-day sessions attended by most lead teachers most times (as shown earlier).

Based on what was reported from observations of the professional development days, it was clear that schools were doing what they felt they could in terms of cascade, with lack of time a frequent concern. Often lead teachers said they outlined lessons and problems they had encountered when teaching lessons and cascade teachers were encouraged to ask for help as they needed. We draw the conclusion that while this is probably a realistic strategy in time-constrained schools to introduce colleagues to a lesson in a 20 minute presentation, it is in itself likely to be insufficient to foster in-depth understanding and engagement.

However, in some cases, ICCAMS lead teachers were reported to authentically engage departments' cascade meetings to focus more on mathematical thinking, considering their teaching styles and student engagement. Lead teachers had sometimes (for example, in School 4) trialled the lessons before the cascade training, creating resources for the cascade teachers. This was perceived as effective and helped with cascade teachers' subsequent planning. In the meetings, the teachers reported they had tried the questions and asked the lead teachers questions about directions to take in the lessons, how they did certain things, and about responses they received from students. These teachers, however, also admitted they sometimes quickly discussed two sets of lessons in an hour meeting, saying that, for each lesson, 'you would like to have at least an hour … and it's not possible in schools'.

School 68 (which scored top on the fidelity measure) ran the cascade training relatively thoroughly, booking in time in departmental meetings which gave them over 30 minutes to discuss the materials for each lesson. School 82 reported that time was set aside in a departmental meeting for lead teachers to feedback their experiences of the training and give an outline of the lessons. Regarding the duration of the cascade training, one teacher said 'it can be ten minutes, 20 minutes …': it depended on how much the lead teacher had amended the lesson. Similar durations were reported by School 81 where teachers said they had 20 minutes in their lunch or break time and said, '[It] should be one hour, but honestly it's never happened.'

Asked about any obstacles faced in the intervention, one lead teacher in School 41 (S041T008, with low fidelity measure as noted in earlier quotes) said, '…definitely the cascading for the other teachers; that has been difficult'. They met during KS3 meeting time but said the cascade became difficult when different teachers were at different stages—it was hard to meet to discuss a particular set of lessons. They then made slides for the cascade teachers that broke down the lessons. One cascade teacher said the lead teachers mentioned how ICCAMS was going, updating on progress and relaying findings, with the leads warning them of difficulties they had with their classes, things they would need to 'scaffold or skip'. Later on in the project, another teacher said the cascade became 'a lot more of "here's the lesson, off you go"'.

One teacher reported that only the lead teacher had a copy of the ICCAMS handbook whereas in other schools each teacher had a copy. Regarding the model of how the material is cascaded between teachers, this teacher suggested that perhaps schools could vary which teachers attend the PD training so others could be engaged more with the project aims.

In School 84, cascade teachers said they have 15 to 20 minutes of sharing the lesson once every half-term. However, they claimed these could go into more depth regarding the potential outcomes of the lesson or possible responses from students. One teacher in this school said they observed an ICCAMS lesson taught by a lead teacher and that this made them more confident in delivering the lesson, allowing them to see where the pupils' thinking was going. This lead teacher also helped 'team teach' ICCAMS lessons with some cascade teachers. So although the cascade training was not extensive, the material and intentions were relayed via observations and discussions. Lead teachers in other schools said they had offered the cascade teachers the chance to observe their lessons but none had taken the opportunity, citing time.

Thus, we conclude, even in more positively reported cases, the school-based PD might be judged to have been inadequate in effecting significant change in teachers' understandings of pedagogy. A caveat to this would be that this does not necessarily mean there could be no change in practice: the lesson materials in themselves can mediate new practices if the teacher's pedagogy is at least amenable to such practices. We noted above, for instance, that some teachers believed that ICCAMS lessons should include 'more discussion', even though the kind of discussion involved might not have been quite as intended.

However, we also conclude that the understanding and practice of cascade was highly variable and confused. This is consistent with our doubt about the effectiveness of the intervention's cascade design and its implementation.

Teachers' recommendations for improvement

Teachers in the regional PD reported that they needed to practise the mathematics of the lessons and think through the lessons' likely learning trajectories beforehand, understanding the problems and anticipating learner responses, in all their aspects, before teaching them. This was said to be effectively done in lead PD sessions when sharing ideas with other lead teachers, but can also sometimes, if more rarely, be achieved in cascade training sessions with other teachers from the same school, under certain conditions. This helps to reveal the issues students will encounter.

This intervention showed how teachers can benefit from re-engaging with mathematics learning in this way, specifically while considering and discussing the difficulties and the misconceptions students may face. In the teacher survey, one teacher summarised this as follows:

> *'[The lessons are] enjoyable to teach once you get used to them. They do need to be well thought through and understood before delivery; that is, there can be a complexity or a nuance that can be missed if this is not done' (School 39, Cascade teacher, male, TY8=Low, FA8=Medium, IC=Medium, ID: S039T011).*

Teachers also reported getting more out of the lessons after doing them more than once. One said: 'It is an iterative thing. When you teach it a second time, you know the pitfalls and those sorts of things.' This suggests that the impact of the ICCAMS pedagogy on the quality of teachers' understanding and practice, and hence the opportunities for learning in their lessons, is one that needs time to grow. We conclude that expecting to see much change in the first year of an intervention may be problematic. We return to this point below.

Similarly, the idea of familiarity summarises the way students perceived the lessons. One common theme, across schools in all regions, was that teachers said pupils took a while to get used to the ICCAMS approach in Year 7 but that they became more comfortable engaging over time. In the long run, they notice the 'brand', and refer to them as an ICCAMS 'kind of lesson', whether positively or negatively. This raises the question as to how a particular school might manage its introduction of a new brand and what the optimal timelines and teacher involvement might be. In the past, the relevant literature suggested development strategies that take account not just of the time for adequate support, but for differences between different teachers: those who 'lead', those who are 'biddable', and those who resist. The corresponding recommendation would be for school leaders to manage the introduction of change to take account of this: prepare the teachers, but also prepare the learners and their families for what is planned.

Some teachers clearly found the ICCAMS approach challenging and a few were reported to actively resist (at one extreme, one teacher conducted the lesson observed in almost complete silence). Teachers needed to have the skills and confidence to run discussions—often considered by mathematics teachers to be a different or difficult skill. They often found it hard to ensure all students were appropriately engaged. Another aspect they found challenging was the timing of lesson parts. At first, teachers found the 'open' lesson structure difficult to manage. Many believed the lessons would take less than half the lesson time of an hour, but the same teachers often enjoyed teaching students to learn through talk, and to solve problems together. More training as a prerequisite for chairing discussions was one recommendation reported. This involves developing a less transmissionist teaching practice for most teachers, and it can give teachers and students the opportunity to discover the cognitive benefits of sharing ideas in various ways. We note that teachers of different subjects tend to have developed more dialogic practices (for example, in the humanities): perhaps some whole-school, cross-curricular PD strategies might serve this cause well.

Usual practice—the control group

Here we consider the practices in the 'business as usual' control schools in response to RQ9.

An email survey was conducted in June 2018 with the control schools, sent to the school contacts used by the project's communications team (leads of departments or of the project management), with 24 school responses. The questionnaire response data does not support the notion that business was necessarily 'as usual'. For example, over half of those responding said they were engaged in the Mastery programme and others were engaged in other mathematics PD projects or programmes. Unfortunately, we do not have data from the intervention schools to compare, except that it was mentioned in a few of our eight field study schools (this issue only became prominent in our thinking after an analysis of the control school survey quite late on). We noted above how some teachers saw some commonalities between ICCAMS and Mastery aims (as well as some differences). In fact, we can claim that from the responses to a question on whether teachers attended other training (further to those specified in the aforementioned school questionnaire) from 500 teachers in intervention and 520 in comparator schools, we got 52 and 62 non-missing responses, respectively (that is, with a response to 'please tell us what other training' you had). From these responses there were only three references to the Mastery programme coming from the surveyed teachers in the comparator group (this is based on combined data from both Year 7 and Year 8 data collections). Thus, this issue did not raise any alarm in the evaluation team.

Unfortunately, as the school-level survey was held later, this was a finding revealed at the analysis stage of the evidence and it was too late to collect systematic data on what might have been 'business as usual' before or alongside ICCAMS with the intervention schools, where our knowledge of their engagement with Mastery is serendipitous and not systematic (that is, we only know that some were). Even though we have concluded that some similar effects to the intervention might be seen in schools in the 'control group, there is no certainty about the Mastery work the schools were engaged in at this time, or how this compares with that previously researched.

The whole-school Embedding Formative Assessment project was a different EEF-funded programme which was deemed during set-up to be closely related to ICCAMS such that schools should ideally not participate in both trials. In order to minimise the number of these schools involved, the EEF provided the delivery team responsible for recruitment with a list of schools taking part in the above programme. Retrospective exploration of this information revealed that nine of the ICCAMS randomised schools were included in this list. In teacher survey responses there were references to formative assessment training from both control and intervention schools.

The following results of the survey show that in this sample of the control schools (1) in addition to the fact that they had signalled their wish to be involved in a PD development, (2) mostly they reported they had engaged in other PD projects, and (3) that we assume they had chosen such PD from an array of possible developments excepting ICCAMS, but mostly engagement with a Mastery programme was mentioned.

Original motivation for participation

In order to understand what was involved in the 'business as usual' control, we wanted to explore the ways in which these schools might have developed their PD. As part of this effort, we considered it useful to explore their original motivation to join ICCAMS. For instance, we thought a commitment to formative assessment might have led a school to develop their practice in other ways not too distant from ICCAMS.

In response to the question, 'What originally motivated the department to be part of the ICCAMS project?', the following responses were received from control schools:

- demands of the new GCSE and its problem solving focus;
- wanted to be involved in education research;
- resources provided;
- improve the teaching and learning within the department with regards to problem solving, algebra, and multiplicative reasoning;
- reward (£1,500); and
- general comments about always wanting to improve.

What was 'business as usual'?

Another question sought evidence of alternative provisions during the course of the evaluation. We particularly asked: 'Since the ICCAMS project started, is there anything you have done to try and meet those goals/address such issues (for example, other approaches, other projects, professional development, Mastery curriculum, etc.)?'

There were lots of different responses here with varying details, often listing ways schools had tried to make improvements for KS3. Six of the 24 (25%) were classed as 'business as usual'—they were not implementing any other programmes. The following responses were also found:

- 12 of the 24 schools mentioned Mastery;
- six mentioned NCETM, and 'Maths Hubs' were mentioned most often; and
- three mentioned emphasis on PD involving problem solving.

Ten schools had designed or chosen their own initiatives, sometimes alongside Mastery or 'maths hub' work. These initiatives and changes were very varied. Here is a short list of different examples of comments from schools:

- Assessment so that '50% of assessment are now open-ended'.
- 'We have stripped out all other aspects of maths from Year 7 and 8 curriculum, to just focus on the number and algebra topics.'
- 'No one initiative. Just general professional development.'
- 'Regular reviews of progress and followed these with corrective teaching to close any gaps that still exist.'
- 'Created our own bank of "basic skills drill" (three-tier differentiated, two-year weekly HL tasks at 20 questions each which address similar topics but see gradual progression).'
- 'Use of "lesson study" style approach.'
- 'We are using a mastery curriculum (the original definition of mastery, not the Shanghai maths approach). Extended time for each topic, mastery testing mid and end of topic with opportunities for re-teaching etc. … We use "knowledge organisers" to help students know which things they need to memorise. Our lesson starters, homework, and cohort assessment strategy is designed to help students remember by providing

> opportunities for spaced practice. We are starting to develop our design of questions and examples to include the idea of procedural variation. The aim is to help our students generalize their learning and also to explicitly teach the features of a problem that are of interest.'

This last comment reminds us that we cannot be sure what it means when a school mentions 'mastery' in its response any more than we can be sure about the homogeneity of what teachers understood by 'the ICCAMS approach'.

We cannot be sure that there was no 'resentful demoralisation' when a school was denied access to a programme it wished to engage with, but 18 out of the 24 responses claimed they were actively engaged in PD activities or programmes other than ICCAMS.

There could be important areas for research in this domain, and perhaps it would be more realistic to name these schools 'comparison' or 'comparator' rather than 'control' or 'business as usual', pending investigation of what actually transpires in these schools.

We understand that the original group of schools volunteered on an 'intention to treat' basis, that is, the schools (or some of their representatives and gatekeepers) all expressed a wish or at least a willingness to be randomly selected into intervention and 'business as usual' control groups contracted to behave in specified ways. We have reported a little information about the 'control/comparator' group of schools from our survey that indicates about a half of these schools (or half of those who responded to the survey) chose to engage in Mastery mathematics professional development programme while others implemented a scattering of other mathematics professional development activities. We know little about the depth of this involvement but we presume (1) that the Mastery programme varies across the country and may be assumed to have had some positive impacts (according to a recent EEF study, these were positive but were not statistically significant) and (2) that these programmes were selected by those schools as appropriate for their school cultures and conditions (on whatever basis).

In a sense, then, 'business as usual' control means that the ICCAMS intervention is being compared to 'alternative PD' activities, ones that are mostly influenced by Mastery mathematics, which probably shares quite a bit in common with the ICCAMS pedagogic approach (as well as some differences), but which have no special emphasis on multiplicative reasoning or algebra.

This highlights concerns with the notion of 'control' in RCT studies of this kind but also the significance of studying what is happening in the control/comparator group: this is widely understood in the literature (Pampaka, Williams and Homer, 2016b, a) but is here exemplified in this case study. Our problems in this case in understanding the effects of the intervention in part relate directly to this general issue.

Explaining the (null and yet) potential impact of ICCAMS on teaching and learning outcomes

The purpose of this section is to establish possible explanations of the evidence provided so far (including also the impact results). We do this via further reflecting on elements of fidelity across cases and a case study of 'impact of ICCAMS on learner outcomes'. The aim of the case study is to suggest or establish possible explanations considered coherent with the literature for the phenomenon in question based solely on the evidence from our (mainly qualitative) data from interviews, observations, and school surveys related to the process evaluation material discussed above. Taking the previous section on themes as indicative of the 'norm' and 'spread' of data about the responses of schools to the ICCAMS intervention, we sum this up and proceed to focus on one school, which appeared to us to be a positive outlier illustrating both the positive effects of the intervention and also its limitations. The purpose here is to indicate some of the conditions in which the intervention might be successful.

At the time this was originally written, the statistical results on impact were not available: the significance of this being that we are not here engaged in post hoc rationalisation of the statistical results but treating the qualitative work outlined here as an independent source of explanations. We believe this strengthens its credibility.

This study, then, aims to provide 'credible' explanations for, or causes of, the various possible outcomes in the light of the IPE, and so ones that might even appear to contradict in some respects the statistical analyses that were to come—that are reported above. We believe this study stands on its own irrespective of final conclusions about the effect sizes of the ICCAMS intervention as evidence of the range of possible and reasonable explanations for what ICCAMS intervention might have achieved in this case. We chose not to rewrite this study in the light of the statistical outcomes, but rather to reconsider it in relation to the report's final conclusions and discussion later.

Revisiting the fidelity measure and regional differences

Figure 10 shows the locations of case study schools on the constructed fidelity measure. Figure 23 presents the ranges of scores on this fidelity measure for the schools in each region. There appear to have been substantial differences between regions, which were the focus of our observations and data discussed above.

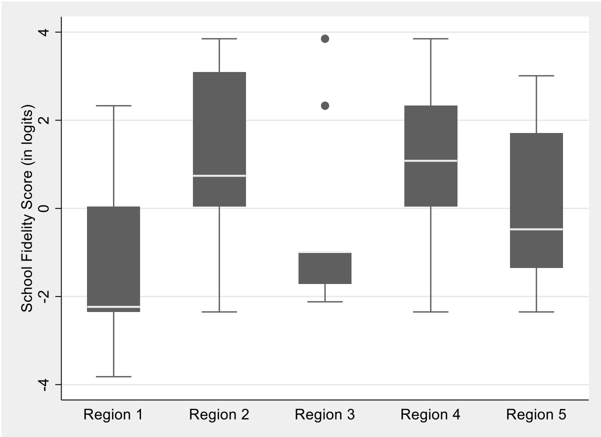
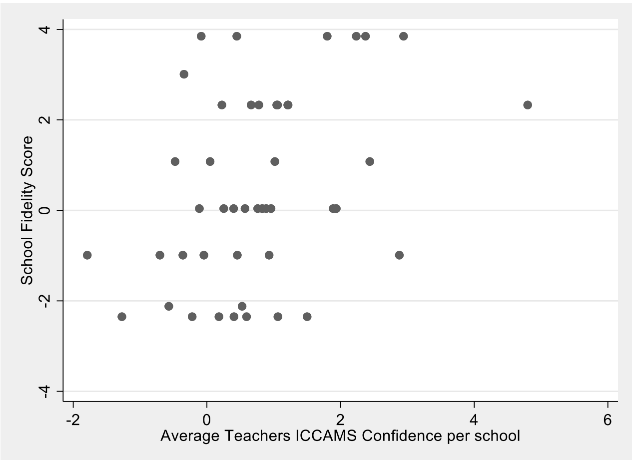


Figure 23: Boxplot of school fidelity measures by region

There is indeed some regional variation in fidelity (in a regression model with school level data, see Appendix 21). Note the point made above that the difference between Region 3 and Region 4 (at the extremes) might be related to significant differences in FSM between these two regions. There is also a positive correlation ($r = 0.38$, $p < 0.01$, $n = 45$, see also left figure below) between fidelity and average (teachers') ICCAMS confidence as also shown in Figure 24.



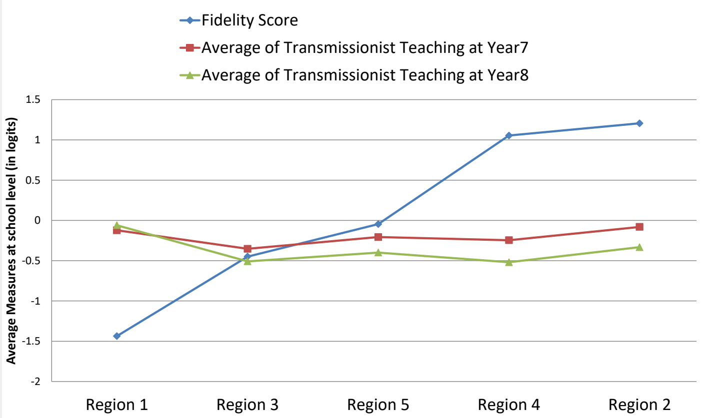


Figure 24: Association of fidelity with other self-report teacher measures (aggregated at school level)

When considering both in the model, their impact is not significant, which might signify an interaction effect. The plot on the right can help understand a bit deeper some of these interactions. It demonstrates how fidelity is higher at regions where there is bigger difference in teaching style (as reported by teachers) between the two years of the study (start of Year 7 and end of Year 8). A more obvious decrease in transmissionism (in Year 8 compared to Year 7) is seen in higher fidelity regions (2 and 4); the lowest fidelity appears in Region 1 where there is no reported difference in teaching style between the scores of teachers in these regions during the course of the intervention. There is, thus, some evidence that engagement with ICCAMS is associated with a change in self-reported teaching practice away from transmissionism and more aligned with self-reported connectionist teaching practices.

The constructed measure of fidelity (see earlier section) uses as items (1) the numbers of ICCAMS lessons taught, (2) the number of LTs attending the regional meetings, and (3) the attendance at cascade training. Table 37 presents a summary of these elements per region and by the fidelity classification (based on the overall measure).

Table 37: Summary of fidelity elements by fidelity classification for each region

	High	Medium	Low		
Overall	15	15	23		
Region 1	1	3	6		
Region 2	4	5	3		
Region 3	1	3	5		
Region 4	5	1	4		
Region 5	4	3	5		
Average of attendance at PD (max 18 for each school)					Average
Overall	16.2	16	13.87		15.13
Region 1	12	14.67	12.67		13.2
Region 2	17.75	16.6	16		16.83
Region 3	16	15	13.6		14.33
Region 4	16	17	13.75		15.2
Region 5	16	17	14.4		15.58
Average of lessons taught (based on all available responses)					
Region 1	18.49	15.43	12.01		15.23
Region 2	20	14.5	12		13.625
Region 3	18.33	16.4	9		15.76
Region 4	17	15.33	7.5		13
Region 5	19.8	14	14.17		17.28
Region 1	17	15	14.17		15.43
Average of cascade (based on minimum rating of lead teachers' responses)					
Overall	2.5	1.92	1.33		1.90
Region 1	3	1.5	1		1.43
Region 2	3	1.8	1.5		2.18
Region 3	3	2	1.5		2
Region 4	2.2	3	2		2.22
Region 5	2	2	1		1.6

ICCAMS cascade PD and the schools' teaching/PD cultures: outliers and the norm

The central question for this subsection is the question RQ7: 'How and to what extent does the method by which training is offered (for example, PD lead or cascade) relate to how ICCAMS is delivered in the classroom?' The visits to schools raised the question of cascade in research data and interpretations more than any other issue—the differences between original 'lead' and cascade trained teachers and their self-reported experience of the PD and teaching being so obvious. Especially concerning was (1) the reaction against ICCAMS by some teachers who said they did not make it work, particularly for 'lower attaining' classes, or because the approach seemed to conflict with their personal beliefs or style and (2) the time it takes for even willing teachers to acquire the skills and confidence to practice ICCAMS lessons or pedagogy in ways they were happy were successful.

A key point here we need to acknowledge: the first year might see a deterioration in some aspects of pedagogy and learner experience, which might be perhaps more effective in following years. On the other hand, there is the potential for a decline in the Hawthorne effect of the experience of just 'doing something different'; this is a methodological challenge we will further reflect on at the end.

Having said this, a broader question perhaps needs to be asked, within which this one about ICCAMS 'delivery' is embedded: how or how much does the teachers' mathematics pedagogy change as a result of the PD experience (evidence is provided on this in relation to transmissionism and FA practice above and with the models of change in students perceptions of teaching practices in Table 32) and, then, what is the likely impact of this on the learners' experience and outcomes, short or long term (we now know the impact short term has been small and negligible on performance measures, but not on their perceptions of the classroom)? This broader question should be in our consideration for several reasons: (1) because the control group *also* were engaged in PD to some degree (for example, many schools chose to engage in the Mastery programme, as detailed earlier) and as a result their pedagogy might also be considered to have been developed, perhaps even in similar ways though perhaps not particularly in relation to multiplicative reasoning or algebra learning compared with mathematics outcomes as a whole, (2) because the teachers' PD concerned general pedagogic knowledge and approach, potentially affecting *all* their teaching and not only the teaching of multiplicative reasoning or algebra, while (3) the experience of the learners perhaps depended *not only* on the developing pedagogy but also on the curriculum materials in use themselves, which served to mediate their learning experience and opportunities.

Points (2) and (3) here clearly interact but it may be that (3) could be expected to explain the experience of learners and outcomes specifically for multiplicative reasoning or algebra more than mathematics as a whole, if a serious difference emerges.

On the other hand, the measured, comparative effects of the intervention (1) might be moderated by any effect of (2) and (3).

The case of School 4

We will now draw on a particular school study that we subsequently characterised as a relatively positive one, recommended as such by the PD lead (as School 4 was one where the in-school development across the department was clearly taken seriously) in the context of the 'norm' found in most schools where cascade training tended to be minimal, and cascade teachers were normally left to make what they could of materials and ask for help when they felt they needed it (see evidence cited above).

Here speaks a lead teacher from School 4:

> *'From my perspective, it's actually made me, it's transformed how I teach. For the concepts and ideas we get from the ICCAMS lessons, I try to bring into my other lessons as well, particularly with the classes that I've been doing ICCAMS with, because we've been quite lucky that we've had the same class in Year 7 and 8 as well, I've really seen a difference in their confidence and their ability to approach problems, they are more willing to have a go whereas before they would have looked at a difficult problem and said "I can't do it" and just given up. …*
>
> *'So, yeah, "3m and m+3"—I really like that … they'll ask questions about it, or you can say, "Have you tried this?''… So that lesson is actually in [the scheme of work] for next year … But I've tried to incorporate questions like that, that require a little deeper thinking, into the mastery' (School 4, Lead teacher, female, TY8=Low, FA8=High, IC=High, ID: S004T015).*

The quote from the lead teacher's interview above connects to most of the themes that we drew out of this case study school: the key phrases and terms being 'transformation', 'other lessons', 'Year 7 and 8', 'confidence to approach problems', 'they'll ask questions', 'actually in the [scheme] for next year', and 'deeper … into the mastery'.

As such, we draw on both qualitative data from observations and interviews to make the case that the school PD culture might vary significantly in explaining differences in classroom activity, and then subsequent survey data that situates the school against the norm regarding school culture and outcomes. To support the discussion, we place the profile of this school against the distribution of all/other (intervention) schools in our sample, considering the average teacher measures as well as the average students learning outcomes.

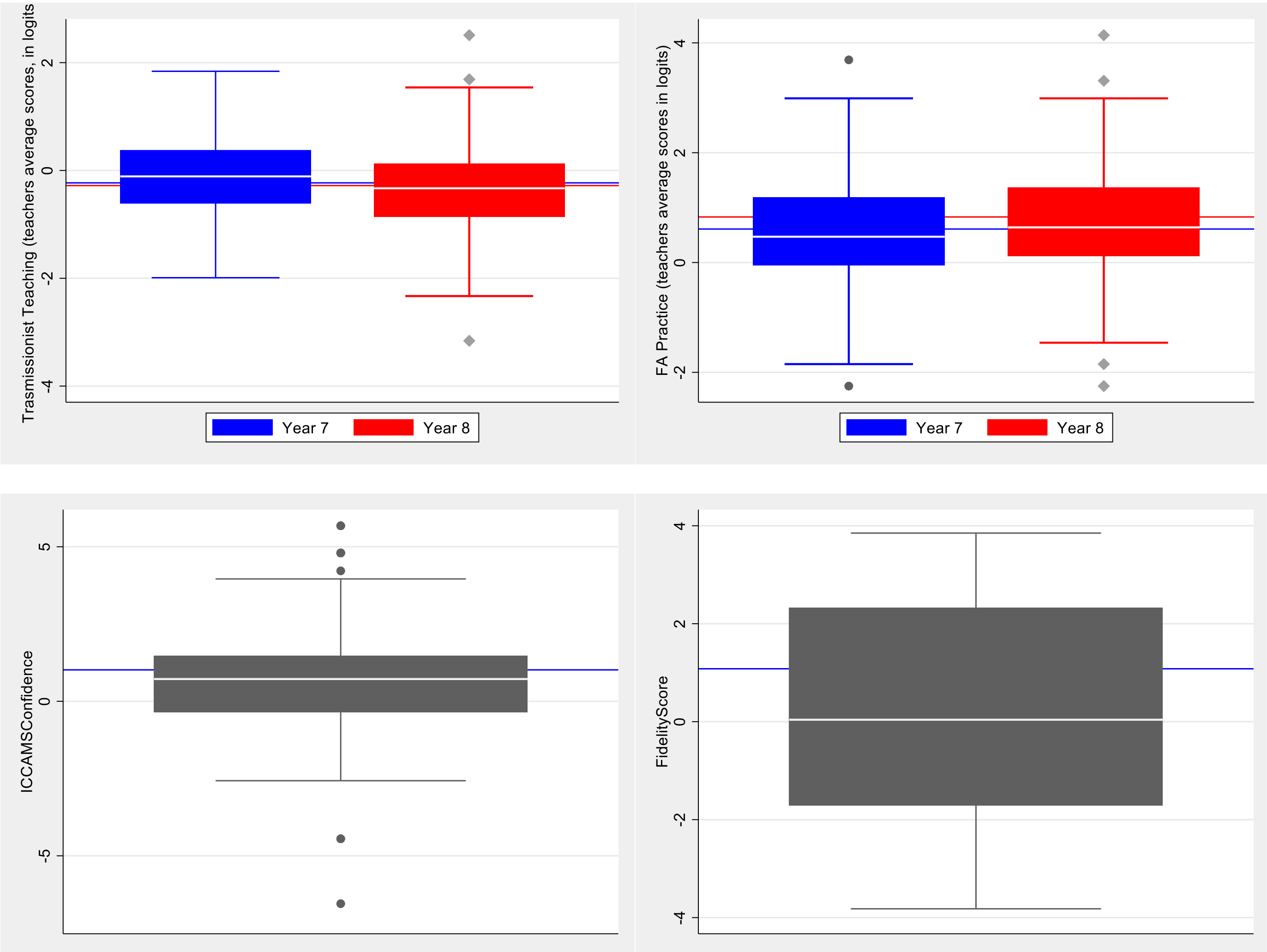


Figure 25: Boxplots of average teacher measures and fidelity scores and the relative position of School 4 (blue and red horizontal lines indicate their corresponding mean for teacher measures or score for fidelity)

From Figure 25 one can observe that the teachers of School 4, on average, score around the median for transmissionist teaching compared with all other intervention schools. Their average scores on FA practice are within the upper quartile, 'medium high', from the median for both Year 7 (0.61) and Year 8 (0.83) and also for confidence in teaching with ICCAMS (1.02 logits). Finally the school's fidelity position is located at the same range (1.08 logits).

Figure 26 is based on average student outcomes for the teachers—only from intervention schools—with School 4 position again shown with lines: mean total score, 16.2; mean algebra score, 4.11; mean multiplication score, 7.03; mean mathematics disposition at DP1, 0.16; and mean mathematics disposition at DP2, -0.16.

In sum this suggests that School 4 is a lower attaining school and with lower mathematics dispositions, but with relatively high fidelity, confidence in ICCAMS, and formative assessment practices, while reporting a relatively lower level of transmissionist pedagogy.

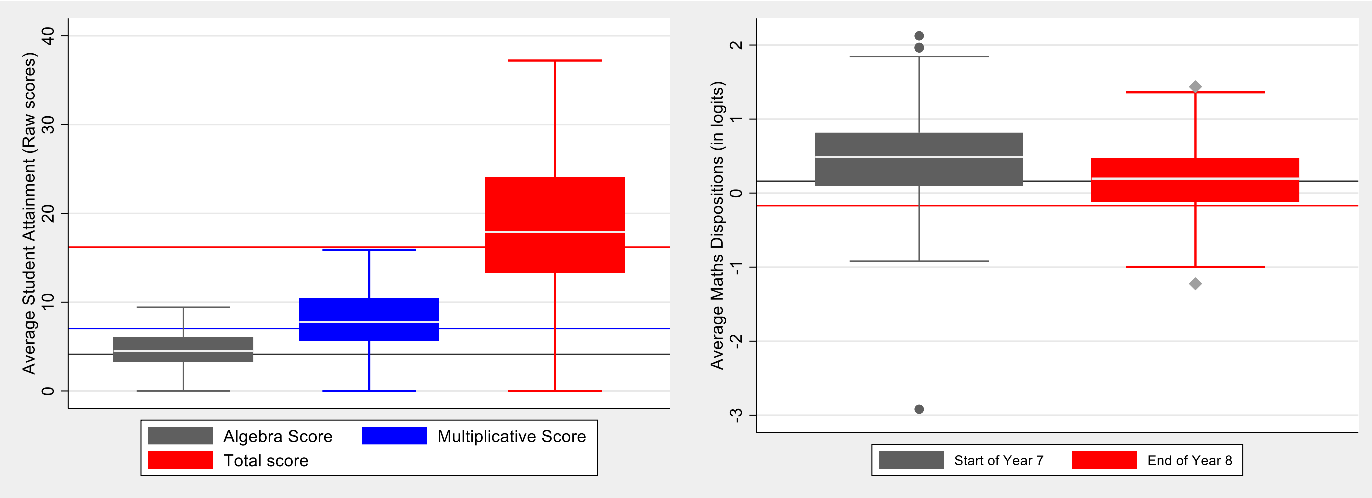


Figure 26: Boxplot of average students' primary and secondary outcomes (aggregated by teachers) and location of School 4

The data from interviews and observations suggested that School 4 was substantially different from the 'norm' in that there was a perception that some pedagogy was transformed and learning was directly affected. This was also evident from the quotation presented earlier from the two lead teachers.[29] One of them said she had been inspired by the regional sessions, which were clearly aligned with the materials: indeed, an observer of one of her lessons wrote, 'This is the most effective ICCAMS lesson I have seen so far.' Her partnered lead teacher adds how important the PD days were, giving them 'the whole day to plan a pair of lessons … which was great because we also got other people's ideas … different ways to teach it'.

Consequently, these two leads confronted their head of department about the needs of cascade and persuaded them to put the cascade of these lessons into the department's scheme of work and planned time in department meetings to work with the cascade teachers, working together on them in school as they had done in the regional PD. The lead teachers, sitting down with their head of department, 'did the meeting plan for the whole year … put in times for ICCAMS … that meant that we definitely had the slots to do it in' on days known well in advance. It also gave the project 'official' recognition that this was now part of the scheme of work.

They made a point of this—believing that in some other schools meetings were squeezed into lunchtimes, which they said is 'not fair'. Finally, the lead teachers appreciated that the first attempts at ICCAMS lessons don't always seem successful: 'It's keeping going … getting them to carry on doing it … doing it regularly, I've really started to enjoy it.'

But there was another strong feature of the way these two lead teachers worked: they decided to plan and teach the lessons together so they observed the student responses to their early attempts and used these to inform the cascade meetings that always followed. They regarded this as important, and invited other teachers to join in their lessons too (actually this had not happened at the time of our observations, but the lead teachers were optimistic that it would happen). Team teaching and observation were mentioned by teachers in some other schools also, and should be recommended as an approach that can be helpful.

Our interview with the teachers at the school revealed another very significant element: the school had been engaged in working on their curriculum and pedagogy via the Mastery programme linked to their NCETM Maths hub (one of the cascade teachers had been deeply involved). One aspect of this was the introduction of what he called a 'pre- and post-test' approach: he said the extra burden of the pre-test was justified because it was designed to inform the teaching; indeed, he said 'to find out what the pupils already know and can do', which almost perfectly reiterates Ausubel's dictum that underpins formative assessment: 'Ascertain what the student knows, and teach accordingly' (Ausubel, 1968).

[29] The lead teachers were interviewed together in this school and it was not always possible to specify who was providing the quote so we do not always quote with a teacher ID in this section. Similarly, we interviewed three cascade teachers in this school but two of them could not be matched with surveys, therefore they are not given IDs either.

The lead teachers claimed that the ICCAMS approach fitted with their Mastery work very well—indeed, this makes clear that the Mastery programme was at work not only in the comparison schools (see below) but also in intervention schools —and helped gain the acceptance of ICCAMS into the department's culture as well as curriculum scheme. One of the cascade teachers who had been heavily involved with the NCETM had written lessons into their Year 7 and 8 curriculum over the previous three or four years, and teachers had been introduced to these lessons in the same manner as the ICCAMS cascade (lessons initially tested and developed by a 'lead' are introduced in department meetings, and then other teachers feedback after trying them out, usually with ongoing discussion informally).

> *'[The Mastery and ICCAMS] are actually closely linked. So [in the Mastery] we have a pre-test and a post-test … The idea behind the pre-test was to inform teachers of the kids what they already know… A bit like the ICCAMS mini assessments' (School 4, Cascade teacher, male).*

Another very significant aspect of their approach to that Mastery programme had been to develop their pedagogy to be responsive to students' own questions, which might be understood as directly engaging with the metacognitive aspect of formative assessment that has been emphasised in the literature: students ask questions when they are framing their own understanding of what they know, do not know, and feel the need to know. The lead teacher quoted previously alludes to this, but the following cascade teacher's remarks make this more explicit:

> *'[In the Mastery maths], the way you are approaching a lesson is drastically different. I suppose the first thing, the biggest thing that was different about it, was the questions coming back at you … that is scary to begin with, you need a more secure knowledge of what you are doing. [But] it's more beneficial for the kids to take on the kinds of questions that they have been coming up with… so for example … "is it 3n or n+3?"—I really like that … they'll ask questions about it' (School 4, Cascade teacher, male).*

Thus, this teacher makes connections between the ICCAMS whole approach and one specific mini assessment and lesson pair ('3n or n+3': 3A, 3B) in regards to their work based in three or four years of development via Mastery, which is about developing the pupils' independence and a more responsive pedagogy led by the pupils' own work and questions. It would seem the experience with Mastery has had a major impact on the way the school has engaged with ICCAMS, even in terms of their culture of working as a department in ways that supported their cascade: a number of interviews by different staff revealed that the staff shared these developments together and were often discussing pedagogy informally.

In sum, this school seemed culturally well prepared to benefit from ICCAMS and their previous history as a department was an important indicator of that, especially their previous work with the NCETM on Mastery. Nevertheless, we could still point to some reservations in the data. The issues of 'time for cascade' and of questions of engagement with some of the materials, especially with 'low attaining classes', were mentioned as problematic issues here also, but teachers were broadly positive about even these two aspects, feeling that these constraints could be overcome over time and with adaptations. The teacher involved most directly with Mastery thought in terms of years to develop such new ways, and this factor seems highly relevant to the evaluation of the impact of the intervention over the two-year timetable here (and commonly in such RCT studies).

Here we summarise the key elements of this school experience and practice that relate very pertinently to the case study as a whole.

1. The lead teachers attended the Region 4 regional PD where attendance was high, the sessions perceived as inspirational, and there was significant evidence of discussion of how to cascade, at least initially.

2. The lead teachers were apparently enthusiastic about ICCAMS, the training, and had good quality outcomes (one of their lead teacher lessons was observed and reported, 'the most effective ICCAMS lesson seen so far', and the teacher concerned reported her teaching had been 'transformed' and that the response of the children was 'amazing').

3. The lead teachers demanded, and got, the support from department for cascading, so they programmed ICCAMS into the whole year's department meetings (although they did not get an hour; they said about 20 minutes but it was informed by their own experiences in trialling the lessons).

4. For each lesson, the two lead teachers taught the lesson together, before the cascade meeting (sometimes more than once).

5. This school had been working on Mastery for several years, led by one of the cascade teachers, and had embedded formative assessment into its programme in Year 7 and claimed it liked the synergy between its implementation of mastery (including, for example, pre-tests for diagnostic purposes) and ICCAMS. The approach was particularly focused on encouraging teaching that addresses students' own questions, implying a metacognitive approach.

6. The attitude to 'low ability' was relatively positive. Although it was admitted by one cascade teacher they might wish to do some adaptation using more scaffolding, they argued that the lessons could still be good for them. This appeared to be part of the new departmental culture, which was said to be previously more traditional.

7. Some teachers had been lucky to teach the same class in Year 7 and Year 8 and felt this important as some teachers new to ICCAMS teaching it in Year 8 were taking time to settle. In general it was suggested the lessons need to be taught several times to get the hang of it; two years was said not to be a great deal of time, though this school was embedding ICCAMS lessons into their scheme for Year 7.

8. One lead teacher lesson early on was critiqued by the lead PD observer as a case where the lesson was unfortunately curtailed, rather than extending into a future lesson: even the lead teachers expressed the need for time for the new practices to bed in.

This is quite strong evidence of overcoming many of the weaknesses in the 'norm' detailed previously and well evidenced in the data coming from the data corpus of the other schools, and it suggests the need for many elements all pulling in the same direction but also supports some of our findings about:

- the potential significance of the Mastery programme in interfering or interacting with ICCAMS and 'control' schools;
- the elements that seemed necessary (if not sufficient) for cascade to be effective;
- the importance of time to embed the ICCAMS practice, mitigating against measurable effects on learners in the short term; and
- some hesitant evidence here that the teachers' adoption of 'Mastery' had focused on a metacognitive pedagogy, which is probably the more demanding aspect of formative assessment pedagogy according to our reading of the literature.

Conclusions to the IPE

We here revisit the research questions that we have addressed in the IPE analyses above, summarising our findings.

We showed substantial variations in the cascade training delivered in school, and in teachers' attendance, with many accounts suggesting only superficial depth in the quality of engagement compared to the regional PD experience (RQ6). On the other hand, the PD leads were effective and used the ICCAMS materials to good effect and this was often noted by lead teachers who attended their regional PD sessions. We did have some evidence, however, that the satisfaction with the preparations for cascade was less than the lead teachers' preparation to teach their own classes, which supported our conclusions about the weakness of the cascade implementation.

Thus, in relation to RQ7, we inferred that the differences between cascade and lead teachers' PD experience was major. Some excellent lessons by lead teachers were observed by lead PDs but we also noted that some cascade teachers' ICCAMS lessons were also observed to be effective. Some teachers, however—usually cascade teachers rather than lead teachers-—expressed concerns that the ICCAMS pedagogy did not suit their classes or their expertise, but one shared concern was teaching ICCAMS lessons with classes of lower attaining students.

The ICCAMS materials were used in cascade PD sessions, by all accounts, but we were told that often this would be done in a brief session where a particular lesson plan would be introduced (20 minutes was the kind of time allocation for this). Consequently, unlike the PD sessions for lead teachers, there was not much room for the kind of dialogues that could serve as a model for classroom pedagogy. We noted many constraints in schools that led to this pressure of time and our case study of one of the more successful schools showed how some of these can be addressed (staff commitment, getting leadership on board, providing more timely PD for staff about to teach particular lessons, team teaching, and so on).

In regards to RQ8, we have noted considerable variation in the engagement of students in ICCAMS lessons, which depended on the teachers' own commitment and pedagogical practice. At base, if teachers were not yet convinced of the ICCAMS approach and activities then the experience could be weak, and such teachers might never establish the confidence to teach that way and even give up on the programme without completing the series of lessons. When committed teachers established new practices consistent with the programme, teachers' and learners' confidence in the approach grew; but such teachers did claim this took some time to bed in.

There was some evidence that 'control' schools had engaged in PD programmes via NCETM hubs that they described as 'Mastery', and other programmes (RQ9). Our team were aware that in some local hubs their Mastery programmes used ICCAMS-type materials and pedagogic approaches (for example, through programmes of multiplicative reasoning) but we did not have enough data to assert more than a possibility about what this involved and how deep it went, nor how it impacted on measured outcomes.

With regard to RQ10, we did note that teachers' previous PD and preferred pedagogical approach can have a major effect on their attitude towards ICCAMS and the way it is implemented in the classroom. This was true for lead as well as cascade teachers.

In toto, this IPE study has offered well-grounded and credibly plausible explanations for effective or non-effective impacts of ICCAMS on teachers and learners, based on mainly qualitative analysis of data from teacher interviews, surveys, and observations.

Importantly, there are key moderating and mediating conditions that might explain why certain relationships in statistical analysis might occur, or alternatively why they might not be as significant as anticipated in the design.

1. The delivery of the ICCAMS professional development was perceived, for the most part, to be effective for lead teachers and in some schools but less so for the cascade teachers (with exceptions like School 4) whose PD, generally, fell far short of the PD the lead teachers experienced.
2. The changes in pedagogy (for example, as measured by their self-reported transmissionism) were observed to be greater for some lead teachers for several reasons—selection, quality of PD, and degree of commitment. The less transmissionist teaching was more associated with effective formative assessment, and so ICCAMS outcomes for attitudes and perceptions of pedagogic practice, and though there may have been no overall improvement in attainment, there have been better outcomes for attitudes that might be assumed to be relevant to attainment.
3. Changes take time and may have less effect, or possibly even negative effects, in the short term on learners' attainment and attitudes in some contexts, especially when the teachers have reservations, resist, or are still embedding the practices into their pedagogy for much of the period of the implementation.
4. The 'business as usual' of the control/comparator group in the context of clustered trials that take place for longer periods such as this two-year intervention is very complex to conceptualise as a genuine control and even more for a randomly chosen school from this 'intention to treat' population: such schools were motivated to change by volunteering to participate in the trial and evaluations of this type will need to collect a lot more information about what their 'business as usual' involved, given this motivation.

Cost

As noted earlier, cost information was based on the costs of training provided by the developer and delivery teams and on the assumption that the intervention is run as intended (that is, two teachers per school attend, and so forth). Data was collected from the developer and delivery teams as well as directly from schools to uncover the expected and any unexpected costs. Based on this evidence provided by some schools and the providers, the cost of ICCAMS was estimated as summarised in Table 38.

Table 38: Cost of delivering ICCAMS

Item	Type of cost	Cost	Total cost over 2 years (per school)	Total cost per pupil per year over 3 years
Fees for services				
PD lead fees for PD delivery (including prep for delivery and delivery)	Start-up cost per school	£41,796.00/ 55 schools	= £759.93	£5.99
PD lead fees for school lesson observations: 2 lessons in each school each year	Start-up cost per school	£43,014.00/ 55	= £782.07	£6.17
PD lead travel to PD and observations costs	Start-up cost per school	£5,960.38/ 55	= £108.37	£0.85
Venue and catering costs (lunch and refreshments provided) 9 sessions in 5 areas: 45 sessions in total	Start-up cost per school	£24,429.29/ 55	= £444.17	£3.51
Purchasing resources				
Materials (including initial handbooks and revisit materials)	Start-up cost per school	£9,315.00/ 55	= £169.36	£1.34
Travel and teacher cover—for two teachers (costs for the intervention group schools in this trial)				
Teacher travel expenses to PD: average of 60 miles return (£0.40 per mile in shared car)	Running cost per school		60 x 9 PD = 540 miles 540 x £0.40 = £216 per school	£1.71
Teacher cover: average of £360 per school per day (£180 per teacher)	Running cost per school		£360 x 9 PD = £3,240 per school	£25.58
Total				
Total (if teacher cover is estimated as £360 per school)	n/a		£5,719.90	£5,719.90/190 pupil average = £15.05 per pupil per year = £30.1 for two years = £45.15 per pupil over three years
Total (if teacher cover is estimated as £180 per school)	n/a		£4,099.90	£4,099.90/190 pupil average = £10.79 per pupil per year = £21.56 for two years = £32.37 for three years

The cost is estimated to be approximately £5,720 over two years for a secondary school, including teacher training costs. The average 'per pupil' cost of the intervention is therefore around £15 per year.

Total training costs (fees for services) are therefore equal to £1,047.27 per school per year. This figure includes background costs for professional development and the visits the PD trainers made to the schools (incorporating fees and travel costs for trainers, venue hire, and subsistence costs for the training).

Costs of purchasing the resources (handbooks and other materials for all teachers) was £169.36 per school over two years.

In this trial, the costs to the schools in the intervention group were teacher travel to PD and teacher cover for PD days. We conducted a short survey of schools to ask about costs. Schools said that little additional printing of resources was required. Two teachers from each school attended professional development—in almost all cases sharing car travel paid per mile by the school, and schools were, on average, 30 miles from the PD training location. Regarding costs of supply teacher cover for those attending PD, we received figures between £180 and £200 per teacher (although some schools provided internal cover and did not suffer this cost).There were no financial costs associated with cascading the material to the department.

There were, on average, 190 pupils per secondary school (20,741 children across the 109 initially recruited schools).

In regards to personnel (teacher) time needed, Table 39 provides an overview.

Table 39: Personnel time for training and preparation

	Lead teacher	Cascade teacher
Training time	9 days (for two teachers each)	9 hours
Preparation time	Up to 1 hour (per week)	Up to 30 minutes (per week)

Overall, 24 days of staff time are required for training: nine days each for two maths teachers that attend professional development training as well as the equivalent of one day (nine cascade sessions of around an hour) for each other maths teacher for in-school cascade training (there were on average eight maths teachers per school). Outside this, the planning time for teachers for ICCAMS should be similar to that of other maths lessons.

Conclusion

Here we integrate the conclusions of the study of impacts on learners and teachers with those of the IPE to formulate fully integrated conclusions of this evaluation as a whole.

Table 40: Key conclusions

Key conclusions
Pupils in the ICCAMS schools made, on average, no additional progress in mathematics compared to pupils in the other schools. This result has a moderate to high security rating.
Exploratory analysis suggests that there is no evidence that ICCAMS improved pupil progress in multiplicative reasoning or improved attitudes to mathematics compared to pupils in other schools but that pupils in schools that received ICCAMS did make the equivalent of one month's progress in algebra.
Pupils eligible for free school meals in ICCAMS schools made the equivalent of one month's progress in mathematics and in the subscales of multiplication and algebra, on average, compared to equivalent pupils eligible for free school meals in the other schools. There was also some evidence of a more positive attitude to mathematics. These results may have lower security than the overall findings because of the smaller number of pupils.
Teacher surveys found that 78% of lead teachers and 54% of cascade teachers said they were confident about ICCAMS teaching. Additionally, student and teacher surveys found some evidence that the intervention did change teachers' practice.
One significant challenge was the cascade training. Only 55% of lead teachers reported managing all the expected cascade training sessions. In addition, although each cascade session was expected to be one hour, only 13% of teachers reported that the sessions were at least this length.

Impact evaluation and IPE integration—implications for future practice and research

In these sections we discuss the evaluation in an attempt to shed light on its design (for example, its logic model) and limitations and hence what we might learn for such work in future. In doing so, we are cognisant of the very strong base of literature that has supported formative assessment while at the same time we have reported mostly null results in this particular intervention study. Hence the imperative of an informed discussion of the particulars of this ICCAMS design and implementation, and the methodology of this evaluation. This is important not just for researchers, evaluators, and intervention designers and implementers in future work, but for teachers and managers considering developing formative assessment practices in their classrooms and schools. They may read the overall null effects together with the positive case study in one school in diverse ways according to their school and classroom conditions, their staffing and in-school training capacity, or even their evaluation of the importance of attitudes compared to attainment outcomes. Professional judgment concerning context—and not only the above empirical findings—are salient in reading this section.

Evidence to support the logic model

If all were to go according to plan, the logic of the intervention would include five main steps (see Figure 4).

1. The design of the lessons (which has been tested in previous research and piloting) matches the curriculum requirements and so the testing of the learning outcomes for the national curriculum for mathematics (including a strong multiplicative reasoning and algebra component); in addition, the lead professional developers (lead PDs) have worked on the design to prepare their programme of support for teachers appropriately in light of the above.

2. Then the selected 'lead teachers' from each school work together in groups with their lead professional developer (there are five lead PDs for approximately 50 interventions schools, two lead teachers in each, that is, 100 lead teachers in total) to ready these lead teachers to teach the ICCAMS lessons to their own classes and to lead their colleagues in their departments (approximately another 100 'cascade' teachers or more) to follow them similarly.

3. The cascade teachers are led in the school PD programme by their own school's lead teachers to understand the lessons and teach their classes appropriately.

4. The result is the ICCAMS lessons are taught appropriately in the classrooms (of circa 200 x 50 = 10,000 children) offering these children a series of ICCAMS tasks and 'learning experiences' including formative assessment, task engagement, dialogue, and reflection and metacognition.

5. This causes better learning, more engagement, and higher achievement or attainment than would otherwise have occurred, which is made visible in tests of mathematics and questionnaires of attitudes.

In practice, we know that such logics can and usually do 'go wrong' at least to some extent at each such step, and our data suggests why and how this sometimes happened, though the data Is not adequate always to judge the scale of this and the significance of this for the impact.

Step 1: The design of the intervention with regard to PD and the plans and guidance in the handbook

The training and guidance was perceived to go well, by all accounts. The lead teachers' reports were ultimately positive of the PD and in at least one case outstandingly so (see the school case study).

One doubt here was about the notions of fidelity versus flexibility (for example, is it *always* desirable for a teacher to rigidly follow a lesson plan made up by the design team? And how far should adaptation be encouraged?). Obviously, this is an issue when the teachers' pedagogic preferences or style contradict the intended approach of the intervention, and here the intervention team and lead PDs hoped at least that such teachers would be supported to 'have a go'. But this issue of adaptability became a particular issue later for some teachers, particularly when working with what we have described as 'lower attainers', which many teachers refer to as 'low ability'. We would expect this to become increasingly an issue over the long term as the pedagogy gets locally embedded but we are not sure that this was much discussed in the ICCAMS development nor subsequently in the regional or school based PD.

There might be some issues with some of the lessons' pedagogical connectedness with the mini assessments, as noted earlier (IPE, "The ICCAMS Teaching material" section): the lack of explicit discussion in the handbook, some teachers' requests for more examples and explanations and the fact that we did not see this issue discussed in PD sessions all suggest that many teachers would need help with this. This pedagogical connection is a subject extensively discussed by Ryan and Williams (2007) and it is not a matter that might be resolved in a brief PD presentation on the topic, a suggestion therefore will be to make this more explicit in the material..

Of course, few teachers will have such multiple experiences within the project period: the alternative is that after early skirmishes with the lessons teachers might start to reflect on the principles of FA, see the value of careful study of the whole set of handbook materials to identify the learning trajectory involved, and that this would mediate their future development. This degree of sophistication in PD needs structured support and might be managed through peer team teaching, lesson study, or even coaching in the early period of the programme.

We conclude that the project could or should have anticipated a longer term development of pedagogic practice and, consequently, that the formative assessment practices might take some years to 'sink into' pedagogy before the intended learning gains become visible. We might, however, expect some impact on teachers' beliefs and practices to become noticeable sooner, that is, in the secondary impact measures of pedagogy.

Step 2: The five PD leads' regional programmes and sessions

Our observations suggested a range of practices. On the whole, teachers were engaged with the training and in the early stages of the trial the attendance was near capacity, although in some regions this declined with time. Teachers' reports were positive about the opportunity to work intensively at lessons (from planning through to post-lesson reflections) and to share ideas in a well-resourced group of experienced practitioners, sometimes in ways not experienced since entering the profession as indicated earlier by the quotation from the lead teacher of School 84.

The regional PD sessions we observed at their best involved modelling the pedagogy the lead PDs hoped the teachers would use in their subsequent classroom and staffroom sessions, as one would hope, through reflection and discussion and suggesting how the teachers could carry this into their own practice.

Our observations (although limited) have one important reservation: we more rarely saw these groups discuss the school cascade training and what difficulties this might involve (we explore later the difficulties that emerged, where we drew on evidence from survey responses and interviews, as discussed earlier).

It would have been interesting to have witnessed some discussion of how to replicate the kind of discussions taking place with the lead PDs when back in school cascade departmental meetings, especially given the critical element of *time* to explore and reflect so frequently mentioned by lead teachers. The learning workload for teachers was significant: not only were they expected to learn how to teach the ICCAMS way, they were expected to pass on this expertise to colleagues in about a fifth of the time they had themselves taken to learn the same skills.

Step 3: Cascading

In all interviews and observations reported above cascade training appears to be a main problem of programme implementation. The quality of the experience of the cascade teachers is almost always reported as less impressive than that of lead teachers: quotes indicate time is short, cascade teachers have less high quality support, are more likely to be led to prioritise other work, and more likely to find the expected practice non-normative. This was largely anticipated by the PD design team and is highlighted frequently in the literature on PD.

As we also noted, this could have been the result of lead teachers' frustrations with providing the cascade training. Based on the combined evidence, a more efficient cascade could have perhaps involved a more realistic dialogue with lead teachers on the extent of their struggles with cascade. The PD leads could have helped their trainees by suggesting practical strategies such as team teaching, cascade teachers observing lead teachers, or even just doing the mathematics problems together as a group. Such strategies could have been implemented and followed up, and might have led to stronger collective understandings and practices of cascade training rather than each school approaching PD in its own way. But these suggestions do depend on investment of resources of time, in critical short supply in schools.

We conclude that we should anticipate that results for the lead teachers' classes outcomes for learner attitudes and performance would generally be better than those of the cascade, but it is not clear how much of this difference is due to the lead teachers being selected, for instance, on the basis of their expertise or confidence in adapting their pedagogy, their being better attuned to the programme's expectations, or being more committed to the programme, and how much it would be due to their superior quality of PD experience.

Step 4: The learning experience

We saw lessons and heard some reports of lessons that we and the lead teachers or lead PDs did not think were well aligned with the intentions of ICCAMS: as mentioned above, even a lesson observed by a lead teacher that was conducted in silence.

As detailed earlier, some teachers said they did not understand what the assessment tasks were for and did not appear to know how to make good use of them. The significance here of the understanding of the teachers for practice is key: sometimes it seemed that an ICCAMS lesson would simply be one where the task sheet provided was used, and sometimes in such cases the teacher said there was not enough work for the students to do whereas those who understood the ICCAMS approach understood how the time for discussion needed to be extended and had strategies to achieve this, for example, via group work, reporting back, and so on.

Some teachers freely admitted that the lessons did not suit their teaching style. This might explain why the degree of transmissionism in cascade teachers was expected (based on the IPE) to be higher than for lead teachers, but also we might expect them to change their practice less over the course of the intervention.

The translation of any new curriculum and pedagogy into classroom practice requires time: even some positively minded teachers said that the first time(s) they tried an ICCAMS lesson it was likely not to go so well, but the positive teachers said that this can be improved with time as they become more confident in the new ways. Clearly this poses problems for less confident teachers or teachers of less amenable classes whose reaction might undermine the teachers' self-confidence. In interviews, some teachers, even lead teachers, often referred to lessons being 'too hard' for their lower attaining classes, and for some teachers this confirmed a notion that the materials do not suit them.

We consider it may be really important for confidence that such teachers see some lessons working before they experience their own relative failure: mutual lesson observation seemed to be a rare occurrence (and of course this raises the cost) but it is one that is encouraged in the NCETM hub Mastery programmes. We were not made aware of how widespread discussions of such strategies were and conclude that we cannot be sure one way or the other about this point (this was also extensively discussed in the case study earlier).

Step 5: The Impact on outcomes

At minimum, we believed the intervention lessons used the tasks and materials in the lesson plans given in the handbook so the learners had opportunities to tackle the problems posed (we could not say this of the mini assessments as some teachers did not use them or understand their role). We know there were teachers who claimed their practice was in the process of transformation, and we were persuaded that the claimed improved practice should have begun to change the classroom experience of the students. Our interviews with students were not successful in eliciting much reflection of the ICCAMS lessons, however, and it may be that the teacher's revised approach to pedagogy is not one that their students perceive as particularly remarkable (this is a known phenomenon: in secondary schools the students' perception of variation in practice between subject teachers may be more noticeable than—or dwarf—changes thought significant by any one subject teacher, see for example Pampaka and Williams, 2016).

Bearing in mind the time-lag involved, we conclude that there might be a positive effect on these students' outcomes for attitude, and so engagement or possibly even attainment in certain conditions, for example, especially for teachers who have low transmissionist teaching practices and for teachers (likely to be lead teachers) who understood rather than instrumentally complied with the intervention lessons such as those observed in one school. From the models in the impact evaluation, however, we found no evidence of improvement on primary or secondary outcomes because of ICCAMS suggesting that most schools did not make a difference in learning outcomes even though they might have changed practice to some extent in the ways intended.

The teachers most positively affected would be likely to transform *all* their teaching practice, as was evidenced by teachers who said they were changing the way they taught right up the school.

> *'And certain lessons we've taken from this and applied to our Key Stage 4 classes' (School 82, Lead teacher, male, ID:S082T004).*

On teacher measures we found weak correlations. The correlations were an anticipated effect of the intervention on teaching practices, but the weakness is all of a piece with the weakness of the intervention effect on student outcomes. We conclude that there is a likely causal connection between the ICCAMS PD and some of the teachers' pedagogy: (1) increased formative assessment practice and (2) decreased transmissionist practice. As there was evidence for the latter being associated with student learning outcomes, this could be interpreted as a mediating (and potential future) impact, but which was not yet manifested as an effect on the learning outcome models.

Overall, the results of the impact evaluation in regards to the effectiveness of the ICCAMS intervention do not provide evidence in support of the original logic model. No significant effect of the ICCAMS intervention was found in relation to the primary and secondary outcomes. Details and explanations about the possible reasons behind this finding were investigated in the IPE thematic analysis and case study above (including the similarity of ICCAMS with other programmes, such as Mastery, used by some control schools); these are discussed in the next section.

There have been, however, further relationships revealed which might potentially mediate or moderate the effectiveness of the intervention such as the consistent negative effect of students' perception of transmissionist teaching on both primary and secondary outcomes, in attainment and dispositions.

Reflecting on ICCAMS from the combined evidence and going back to the literature, it may have been too soon to observe the effect of the intervention on learning outcomes: interventions created by developers and implemented by others need time to develop in school practice and necessarily end up on the ground in practice as different from the inventor's imagination (as Stenhouse (1975) argued, every new group of teachers becoming involved in a project must develop it to make it their own, in their own way). All departments have their own ways of working and teachers each have their own past experiences and practices. In this particular intervention, the provision of cascade training and the engagement of cascade teachers were both dependant on the time schools managed to give, as were the number of ICCAMS lessons at the standard intended.

The ICCAMS intervention is itself intended to be developmental for students, but also for teachers. It is an intervention primarily with teachers' praxis in focus but one whose primary impact is here being primarily measured by what happens with students. But the gains for teachers are as important because the student experiences and outcomes cannot be understood without those of the teachers, and even if the benefits to learners are negligible in the short time of the project, positive changes in teaching might lead to longer term benefits for learning.

The following improvements are therefore suggested, in brief:

- teachers need to practise the mathematics themselves before teaching, ideally with other teachers preparing for the same lessons, and with the input of other teachers who have had experience of teaching the lessons;

- encouraging observation of, as well as by, lead teachers and joint teaching or lesson study of lessons, and other alternative forms of cascade training that involve joint peer activity and reflection on the new approach, materials, and students' responses;

- teachers asked for advice on differentiation and adaptation especially for lower attaining pupils; such adaptations could be widely discussed and made available, with commentaries—again, especially in order to help low attaining or less confident students and their teachers; more work in PD and cascade should emphasise ways of teaching ICCAMS and FA to these students;

- teachers need help in facilitating pupils' discussions and managing off-task talk: this could be a focus for some parts of the PD;

- some consideration to managing learners' expectations of change could be helpful in implementing new practices ; and

- above all, schools need to give PD more time to achieve changes in teaching and classroom implementation.

A revised and simplified version of the logic model based on the findings is shown below. We have evidence of perceived change in teaching practices and perhaps their mediation of student outcomes; this is presented here as a hypothetical or potential future impact (of course we do not claim this is a finding of this evaluation). Projections beyond the project, based on known findings from other studies in relation to the effects of teaching practices on learning outcomes as well as the reported association of such practices with improvements in average students' mathematics dispositions, could support the effectiveness of this PD programme at least for the lead teachers, as will be discussed below.

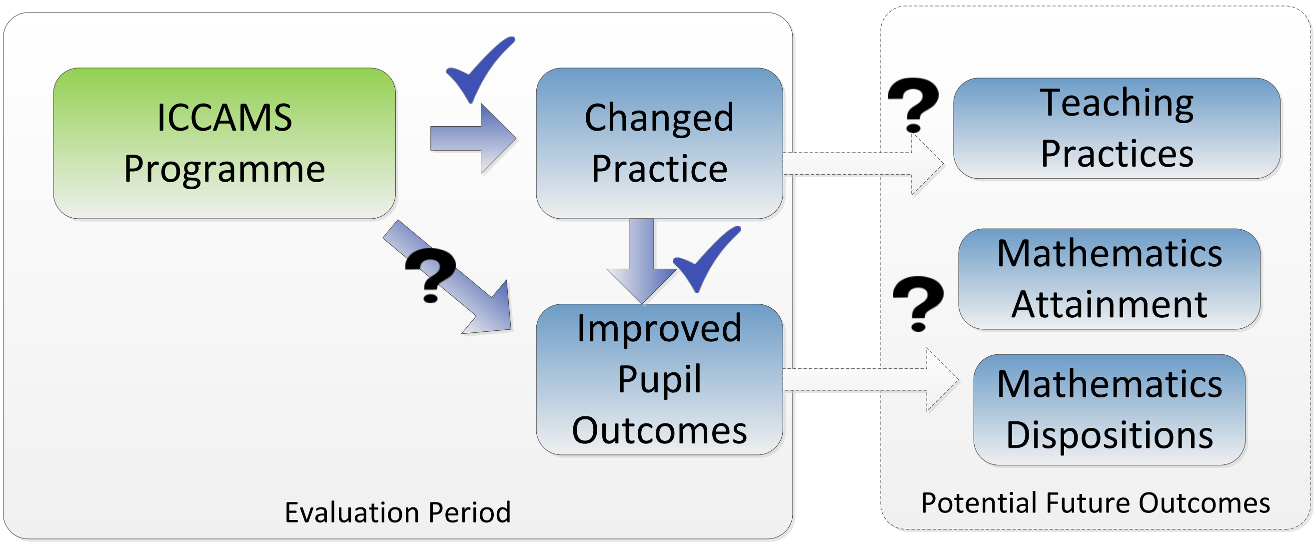


Figure 27: A simplified logic model (left) based on evaluation results with potential future impact paths (right)

Interpretations for future research and the practice of formative assessment

The consideration of coherence of this evaluation with the literature is important to its evaluation and the credibility of any explanations. This is because ICCAMS was informed by research on formative assessment that has strong empirical and conceptual grounds. To summarise this briefly: the dialogue between learners and teachers can provide, under certain conditions, a foundation for engaging teaching and learning practices likely to mediate good, or at least improved, educational outcomes. The dialogue can be viewed in two perspectives. On the one hand, the dialogue informs the teacher about what the learner knows related to the curriculum and its intended outcomes thus informing the teacher and allowing teaching to target the tasks better. In Ausubel's famous dictum, if one were to sum up the implications for teaching of all educational psychology, it comes to this: that what the learner can learn is determined by what they already know, 'ascertain this and teach accordingly'. However, there is perhaps an equally, if not more important, aspect to the dialogue and that is the engagement of the learners' metacognition whereby the student comes to understand what it is that they know or do not know and understand thus informing their own understanding of their needs and what tasks might need to come next (Wiliam, 2007b).

The mini assessments designed by the ICCAMS team clearly attempted to provide tools to elicit learners' knowledge and understandings, and the materials in general aimed to develop significant dialogue around precisely these foci, in important mathematical contexts. Let us take the example of the lesson pair Algebra 3A (Boat hire)/3B(Balloon); the mini assessment asks, 'Which is larger, 3n or n+3?' (see the handbook's Figure 3). This lesson was focal in one of the lead PD sessions we observed and these comments benefit from that videoed observation (see Figure 2). This is an excellent mini-assessment that connects closely with longstanding research (CSMS and ever since) into learners' conceptions of variable. The 'boat hire' lesson has been the subject of lesson studies by the evaluation team's colleagues and students (including two of the present authors) for many years and we have acquired through collective PD substantial experience of it. A recent lesson study report has been published in Archer et al. (2021) for this particular lesson. Consider two scenarios: (1) the class splits in their answers: some say '3n', others say 'n+3 is larger', and perhaps some say 'you can't be sure' or 'I don't know'; (2) the vast majority or all of the class think 3n is larger, perhaps because 'multiplication makes things bigger'. How will the teacher use this information in lesson planning, and in the lesson 3A in particular? This is not made clear in the handbook and we think many teachers would need help with making this connection, as also noted in some of the teacher comments and suggestions for improvement of the materials (e.g. in Table 34). Such an issue might then need to be picked up in the PD both in regional and in school session discussions, but also in reflections on practising the lesson by the various teachers in PD dialogues, cascade, lead teacher, and lead PD. This dialogue might go to the heart of the FA design but engage with the classroom practice in ways that can bring new pedagogic understandings: in short, deep PCK in practice.

Then, there is the key metacognitive element of formative assessment for the learners; we judge that this could be made stronger. In the handbook's general guidance, a teacher might find the FAQs attractive where one reads that the mini assessment tasks should be used before teaching either of the linked lessons, and that 'they have been designed to allow you time to consider how the students are likely to engage with the lesson and what they might find challenging'. This could be a lost opportunity, when an emphasis on the learners' metacognitive experiences required for optimal formative assessment practice could involve coming back to the pupils' responses to the mini assessment after the lesson pair and discuss with them, "How has your answer to this assessment changed now, and why has it changed? What have you learnt from this lesson pair?"' Nevertheless, based on our observations there is no doubt the teachers can benefit from the initial use of the mini assessments in general to plan their lessons, and plan specific questions for discussions they see as likely to arise. But according to the feedback from some teachers many of them might need help with capitalising on this aspect and help in the specific contexts of these lessons. Again, these might be thought subtle points that can best come through the PD.

In the example above, the mini assessment question does link directly to the task in lesson 3B but it does not link directly with the task in 3A, which essentially asks for a comparison of 10+n and 5n. Teachers might be perplexed as to why the lesson pair is designed this way and how this could contribute to the practice of FA. As such links are not directly apparent at the teaching material/text (as also noted in teacher surveys and interviews) it may be useful to ensure that this part of the design is covered and discussed at the PD sessions.

In sum, we recommend that the ICCAMS approach could more strongly and explicitly promote metacognition, which the literature suggests might be the key to improved outcomes from FA.

Reflecting on the efficiency of cascade PD and the challenges, there may be lessons to be learned from the literature and, in particular, recent studies of PD that capitalise on internet technologies and online tools to support and offer complementary alternatives to face-to-face PD activities. Such technologies are considered to be very helpful in bringing groups of teachers together, especially where there are limitations due to geography and resources (Meletiou-Mavrotheris, Mavrou, Stylianou, Mavromoustakos and Christou, 2014; Li and Qi, 2011; Forrester, Motteram and Bangxiang, 2006; Kilde and Gonzales, 2015; Meyers, Molefe, Brandt, Zhu and Dhillon, 2016). Although communities in the U.K. are probably mostly not as remote as in other parts of the world, internet technologies have also been successfully used for PD. For example, Bevins, Jordan and Perry (2011) encouraged teachers to initiate their own small-scale action research projects including the use of STEM (Science Technology Engineering Mathematics) to raise student engagement and motivation in mathematics and literacy. Teachers were encouraged to engage in reflective discussions, with various approaches including audio reflections, a paper-based learning and evaluation tool, and an online hub, which also enabled sharing knowledge gained through their classroom-based research. Thinking about the future, it is likely that internet technologies will become much more popular for PD activities and, if ICCAMS were in its conceptualisation phase now, perhaps these technologies would have been included in the programme. (Reducing our carbon footprint might also now be a consideration.) We therefore tentatively recommend that adoption of the ICCAMS PD approach might benefit from such technology-mediated teaching communities.

There are also sound reasons for developing pedagogic practices specific to key ‘conceptual fields’ that connect knowledge, facts, and procedures in coherent ways. This demands attention to *particular* concepts and the discussion of alternative conceptions—sometimes misconceptions—and mathematical methods, as well as particular pedagogical content knowledge, involving particular models and problem contexts that beg to be organised by the particular field involved. The multiplicative conceptual field is the best researched and most well-known of these ‘fields’ in the literature in the psychology of mathematics education and in the professional development literature specifically related to the Freudenthal school whose experiences have informed some of the ICCAMS tasks.

Finally, it is known that this field of multiplicative reasoning (also including algebra) engages with the majority of the curriculum in at least secondary school up to GCSE level. In short, if learners master this field, their exam results should be very high and the extent to which they master this field should determine the extent of their success in tests up to GCSE. One might argue that this is one area where the demands of performativity actually align with what many have argued is good mathematical understanding and practice.

Regarding this particular ICCAMS programme and plans of activities for lessons, a pedagogical approach is built into the plans and ‘professional development’ that invites more ‘connectionist teaching’ in which children’s own mathematical problem solving is encouraged and discussed. This pedagogical tradition has a long history and is encouraged in much PD work nationally and internationally. In addition to this there is a focus on dialogue, however, teachers find this ‘dialogic pedagogy’ much more challenging than traditional ‘delivery’ or ‘transmissionist’ pedagogies, sometimes also called ‘direct teaching’—challenging both intellectually but also in terms of classroom management. The narrative above reflects this wider evidence from the literature: from experiences of the Numeracy Strategy in England to those of the Japanese Lesson Study around the world, and even Japan, all report that the parts of lessons where teachers engage the class in a dialogue about the various approaches or diverse conceptions surfaced is very challenging for the teacher.

Nevertheless, the advocacy of formative assessment continues as one of the strongest and most likely claims to improving learning outcomes in relevant studies in mathematics education (for example, Wiliam, 2007) and the essential argument is that traditional teaching is largely unsuccessful in developing mathematical concepts and problem solving capabilities, and even more, it robs the learner of the control that in the long run is so important to learner autonomy and metacognition: they develop ‘surface’ rather than ‘deep’ learning strategies (Marton and Booth, 1997; Marton and Säljö, 1976).

The core explanation for any impact of formative assessment in general and ICCAMS’ particular materials and approach on pedagogy or learning outcomes, then, rests on this combination of formative assessment and ‘connectionist’ pedagogic ideas—that better learner outcomes should follow a formative-assessment-led, dialogic pedagogy that elicits children’s own mathematical conceptions and ideas in classroom discussion of the Multiplicative Conceptual Field (MCF) and algebra. The inference is obvious, then: problems in the PD that lead to inadequate understandings and practices in this regard are likely as not to miss the point and so achieve only poor results. And in the test of impact of the intervention compared to a control/comparator group, the question we would have to continue to ask is:

> *How did the intervention play out in terms of the learning-teaching experience of the intervention classrooms in relation to that of the control/comparator group of classrooms and pupils/teachers?*

Our methodology did not allow us to explore the classrooms in the control schools and so we have no definitive answer to this question (we do have, however, comparative perceptions of teaching practice from student and teacher surveys).

A note here is worth making about the generalizability of these findings, especially from the impact evaluation: the sample of schools taking part in this evaluation do not appear to differ from the national average in any significant way and the resulting power of the study was higher than initially planned. Therefore, we consider the results generalizable, assuming repeatable conditions (however, also with the understanding that the reality is much more complex than what is captured in our already complicated analytical models).

Limitations and lessons learned for methodology

In this section we overview limitations and lessons learned from this evaluation that could help the design of future evaluations.

The main limitations concern the following:

- the measurement of fidelity;
- logistics and access to schools to enable more, and more useful, student interviews and direct discussions after lessons; and
- more information on what other developments or initiatives happen in intervention schools as well as more intensive case studies on control schools.

Other limitations naturally include response rates to some teacher surveys (and matching teachers' responses across the two time points), control group surveys, and, to a much lesser extent, missing data from students. In addition, the timing of the intervention and the involved target group (students at the start of Year 7 when randomisation took place) along with missing information from schools makes it challenging to have a true estimate of the student baseline sample.

There are, however, various lessons learned and reflections on this process, which can inform similar evaluations.

Measuring the fidelity of such a complex intervention with three items is not enough to capture the complexity and variation in these schools. Even after agreement of the scoring for the three main elements with the other teams during analysis it was obvious that this scoring was not reflecting practical realities, especially in regards to the quality of cascade and of lessons taught in the school. The amendments we performed (detailed in Table 12) led to the construction of a defensibly reliable measure to capture some broad elements of the intervention but still does not capture all essential aspects.

There is also a point to be made about the importance of considering the class and school level in the design and data collection of interventions of this type: our results showed low variance at school level but much higher class and/or teacher levels, and recent methodological guidelines (Demack, 2019) point to the need to take into account these levels, especially for mathematics in secondary education.

Another key point concerns the timeframes of evaluations: the first year of such an intervention might see a deterioration in some aspects of pedagogy and learner experience (for example, we noted the initial caution or 'resistance' of many teachers and how some were resistant to change in the longer term) while becoming perhaps more effective in following years. This limits the effectiveness of evaluation studies that attempt to show results in the short term, that is, evaluating PD impact by measuring immediate impact on learner outcomes. Rather, future trials should look to the change in teachers' pedagogic knowledge first, understanding beliefs and practices in relation to the variation in PD experiences, before examining how these translate into learners' attainment in the longer term. Even the PD effect on teaching should also be considered over the long as well as short term.

And finally, there may be a lesson to be learnt in relation to the comparator group ('control') in such interventions. There could be important areas for research in this domain. In the attempt to capture 'business as usual', perhaps 'comparative' school or setting might be a more realistic conception and support the kind of work needed to capture what is going on in these schools. This raises questions about the comparative group of schools and whether they provide an appropriate counterfactual within realistic complexities and dynamically evolving conditions the schools operate in. In any case, we reflect in a design such as this, much more needs to be known about the comparison samples.

References

Adey (2006) 'A Model for the Professional Development of Teachers of Thinking', *Thinking Skills and Creativity*, 1, pp. 49–56.

Adey, Hewitt, Hewitt and Landau (2004) *The Professional Development of Teachers: Practice and Theory*, Dordrecht: Kluwer.

Adey and Shayer (1994) *Really Raising Standards: Cognitive Intervention and Academic Achievement*, London: Routledge.

Akiba (2012) 'Professional Learning Activities in Context: A Statewide Survey of Middle School Mathematics Teachers', *Education Policy Analysis Archives*, (20), pp. 1–36.

Angrist (2006) 'Instrumental Variables Methods in Experimental Criminological Research: What, Why and How', *Journal of Experimental Criminology*, 2, pp. 23–44.

Angrist and Imbens (1995) 'Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity', *Journal of the American Statistical Association*, (90), pp. 431–442.

Angrist, Imbens and Rubin (1996) 'Identification of Causal Effects Using Instrumental Variables', *Journal of the American Statistical Association*, 91, pp. 444–455.

Archer, Morgan and Swanson (2021) *Understanding Lesson Study for Mathematics: A Practical Guide for Improving Teaching and Learning*, Abingdon: Routledge.

Ardron and Monahan (2010) 'Assessing Pupils' Progress: Keeping It Real in the Primary Classroom', *Primary Science*, (115), pp. 9–13.

Argentin, Pennisi, Vidoni, Abbiati and Caputo (2014) 'Trying to Raise (Low) Math Achievement and to Promote (Rigorous) Policy Evaluation in Italy: Evidence from a Large-Scale Randomized Trial', *Evaluation Review*, 38, pp. 99–132.

Askew, Brown, Rhodes, Johnson and Wiliam (1997a) 'Effective Teachers of Numeracy' (final report), London, King's College.

Askew, Brown, Rhodes, Wiliam and Johnson (1997b) 'The Contribution of Professional Development to Effectiveness in the Teaching of Numeracy', *Teacher Development*, 1, pp. 335–356.

Ausubel (1968) *Educational Psychology: A Cognitive View*, San Francisco, CA: Holt, Rinehart and Winston.

Baker, Gersten, Dimino and Griffiths (2004) 'The Sustained Use of Research-Based Instructional Practice: A Case Study of Peer-Assisted Learning Strategies in Mathematics', *Remedial and Special Education*, 25, pp. 5–24.

Basma and Savage (2018) 'Teacher Professional Development and Student Literacy Growth: A Systematic Review and Meta-Analysis', https://discovery.ucl.ac.uk/id/eprint/10041629/1/savage_article.pdf

Bennett (2011) 'Formative Assessment: A Critical Review', *Assessment in Education: Principles, Policy and Practice*, 18, pp. 5–25.

Bevins, Jordan and Perry (2011) 'Reflecting on Professional Development', *Educational Action Research*, 19, pp. 399–411.

Black, Harrison, Hodgen, Marshall and Serret (2011) 'Can Teachers' Summative Assessments Produce Dependable Results and Also Enhance Classroom Learning?', *Assessment in Education: Principles, Policy and Practice*, 18, pp. 451–469.

Black and Wiliam (1998a) 'Assessment and Classroom Learning', *Assessment in Education*, 5, pp. 7–74.

Black and Wiliam (1998b) 'Inside the Black Box: Raising Standards Through Classroom Assessment', London: King's College.

Blömeke and Olsen (2019) 'Consistency of Results Regarding Teacher Effects Across Subjects, School Levels, Outcomes and Countries', *Teaching and Teacher Education*, 77, pp. 170–182.

Bond and Fox (2007) *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd edn), NJ: Lawrence Erlbaum Associates Inc.

Boylan, Demack, Willis, Stevens, Adams and Verrier (2015) 'Multiplicative Reasoning Professional Development Programme: Evaluation', Department for Education. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/436835/RB406_-_Multiplicative_Reasoning_Professional_Development_Progeramme.pdf

Burroughs, Gardner, Lee, Guo, Touitou, Jansen and Schmidt (2019) 'A Review of the Literature on Teacher Effectiveness and Student Outcomes', in *Teaching for Excellence and Equity*, London: SpringerOpen.

Cajkler, Wood, Norton and Pedder (2014) 'Lesson Study As a Vehicle for Collaborative Teacher Learning in a Secondary School' *Professional Development in Education*, 40, pp. 511–529.

Carlson (2002) 'A Community of Practice: Web Portals and Faculty Development', *Journal of Computing in Higher Education*, 13, pp. 4–24.

Carpenter, Fennema, Franke, Levi and Empson (1999) *Children's Mathematics: Cognitively Guided Instruction*, Portsmouth, NH: Heinemann.

Chirinda and Barmby (2017) 'The Development of a Professional Development Intervention for Mathematical Problem-Solving Pedagogy in a Localised Context', *Pythagoras*, 38.

Chval, Abell, Pareja, Musikul and Ritzka (2008) 'Science and Mathematics Teachers' Experiences, Needs, and Expectations Regarding Professional Development', *Eurasia Journal of Mathematics, Science and Technology Education*, 4, pp. 31–43.

Copur-Gencturk and Papakonstantinou (2016) 'Sustainable Changes in Teacher Practices: A Longitudinal Analysis of the Classroom Practices of High School Mathematics Teachers', *Journal of Mathematics Teacher Education*, 19, pp. 575–594.

Cordingley, Bell, Evans and Firth (2005) 'The Impact of Collaborative CPD on Classroom Teaching and Learning. Review: What Do Teacher Impact Data Tell Us About Collaborative CPD?', Research Evidence in Education Library, London, EPPI-Centre, Social Science Research Unit, Institute of Education, University of London.

Cordingley, Higgins, Greany, Buckler, Coles-Jordan, Crisp, Saunders and Coe (2015) 'Developing Great Teaching: Lessons from the International Reviews Into Effective Professional Development', Teacher Development Trust.

Darling-Hammond, Hyler and Gardner (2017) 'Effective Teacher Professional Development', Palo Alto, CA: Learning Policy Institute. https://learningpolicyinstitute.org/sites/default/files/product-files/Effective_Teacher_Professional_Development_REPORT.pdf

Darragh and Radovic (2019) '"To Tia with Love": Chilean Mathematics Teacher Identities After Professional Development', *ZDM - Mathematics Education*, 51, pp. 517–527.

Davis, Kumtepe and Aydeniz (2007) 'Fostering Continuous Improvement and Learning Through Peer Assessment: Part of an Integral Model of Assessment', *Educational Assessment*, 12, pp. 113–135.

Demack (2019) 'Does the Classroom Level Matter in the Design of Educational Trials? A Theoretical and Empirical Review', Project Report, London: Education Endowment Foundation.

Department for Education (2016) 'Standard for Teachers' Professional Development: Implementation Guidance for School Leaders, Teachers, and Organisations that Offer Professional Development for Teachers', Crown Copyright DFE-00167–2016.

Desimone, Smith and Phillips (2013) 'Linking Student Achievement Growth to Professional Development Participation and Changes in Instruction: A Longitudinal Study of Elementary Students and Teachers in Title I Schools', *Teachers College Record*, 115.

Ding and Harskamp (2011) 'Collaboration and Peer Tutoring in Chemistry Laboratory Education', *International Journal of Science Education*, 33, pp. 839–863.

Donner and Klar (1996) 'Statistical Considerations in the Design and Analysis of Community Intervention Trials', *The Journal of Clinical Epidemiology*, 49.

Donner and Klar (2000) *Design and Analysis of Cluster Randomization Trials in Health Research*, London: Arnold.

Fennema, Carpenter, Franke, Levi, Jacobs and Empson (1996) 'A Longitudinal Study of Learning to Use Children's Thinking in Mathematics Instruction', *Journal for Research in Mathematics Education*, pp. 403–434.

Fernandez and Yoshida (2004) *Lesson Study: A Japanese Approach to Improving Mathematics Teaching and Learning*, New York: Routledge, Taylor and Francis.

Finau, Treagust, Won and Chandrasegaran (2018) 'Effects of a Mathematics Cognitive Acceleration Program on Student Achievement and Motivation', *International Journal of Science and Mathematics Education*, 16, pp. 183–202.

Fletcher-Wood and Zuccollo (2020) 'The Effects of High-Quality Professional Development on Teachers and Students: A Rapid Review and Meta-Analysis', Wellcome Trust. https://epi.org.uk/wp-content/uploads/2020/02/EPI-Wellcome_CPD-Review__2020.pdf

Forrester, Motteram and Bangxiang (2006) 'Transforming Chinese Teachers' Thinking, Learning and Understanding Via E-Learning', *Journal of Education for Teaching*, 32, pp. 197–212.

Foster, Toma and Troske (2013) 'Does Teacher Professional Development Improve Math and Science Outcomes and Is It Cost Effective', *Journal of Education Finance*, 38, pp. 255–275.

Franke, Webb, Chan, Ing, Freund and Battey (2009) 'Teacher Questioning to Elicit Students' Mathematical Thinking in Elementary School Classrooms', *Journal of Teacher Education*, 60, pp. 380–392.

Freudenthal (1978) *Weeding and Sowing: Preface to a Science of Mathematical Education*, Dordrecht, Netherlands: Kluwer Academic.

Gerber and Green (2012) *Field Experiments: Design, Analysis and Interpretation*, New York: WW Norton.

Gersten, Taylor, Keys, Rolfhus and Newman-Gonchar (2014) 'Summary of Research on the Effectiveness of Math Professional Development Approaches' (Rel 2014–010), Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

Goos, Dole and Geiger (2011) 'Improving Numeracy Education in Rural Schools: A Professional Development Approach', *Mathematics Education Research Journal*, 23, pp. 129–148.

Goulding (2002) 'Cognitive Acceleration in Mathematics Education: Teachers' Views', *Evaluation and Research in Education*, 16, pp. 104–119.

Gravemeijer (1999) 'How Emergent Models May Foster the Constitution of Formal Mathematics', *Mathematical Thinking and Learning*, 1, pp. 155–177.

Gravemeijer, McClain and Stephan (1999) 'Supporting Students' Construction of Increasingly Sophisticated Ways of Reasoning Through Problem Solving', in Olivier and Newstead (eds), *Proceedings of the Twenty-Second Annual Meeting of the International Group for the Psychology of Mathematics Education*, Stellenbosch, South Africa.

Gresalfi and Cobb (2011) 'Negotiating Identities for Mathematics Teaching in the Context of Professional Development', *Journal for Research in Mathematics Education*, 42, pp. 270–304.

Gupta (2011) 'Intention-to-Treat Concept: A Review', *Perspectives in Clinical Research*, 2, pp. 109–112.

Haas (2005) 'Teaching Methods for Secondary Algebra: A Meta-Analysis of Findings', NASSP Bulletin, 89, pp. 24–46.

Hart (ed.) (1981) *Children's Understanding of Mathematics: 11–16*, London: John Murray.

Hatch and Lee (2010) 'Professional Development', in Wilder, Lee and Pimm (eds), *Learning to Teach Mathematics in the Secondary School: A Companion to School Experience* (3rd edn), Abingdon: Routledge.

Hedges and Hedberg (2007) 'Intraclass Correlation Values for Planning Group-Randomized Trials in Education', *Educational Evaluation and Policy Analysis*, 29, pp. 60–87.

Herman, Osmundson and Silver (2010) 'Capturing Quality in Formative Assessment Practice: Measurement Challenges' (CRESST Report 770), Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Heyd-Metzuyanim, Munter and Greeno (2018) 'Conflicting Frames: A Case of Misalignment Between Professional Development Efforts and a Teacher's Practice in a High School Mathematics Classroom', *Educational Studies in Mathematics*, 97, pp. 21–37.

Hilton and Hilton (2019) 'Primary School Teachers Implementing Structured Mathematics Interventions to Promote Their Mathematics Knowledge for Teaching Proportional Reasoning', *Journal of Mathematics Teacher Education*, 22, pp. 545–574.

Hodgen, Brown, Kutchemann and Coe (2010) 'Mathematical Attainment of English Secondary School Students: A 30-Year Comparison', paper presented at the British Educational Research Association (BERA) Annual Conference, University of Warwick.

Hodgen, Coe, Brown and Kuchemann (2014) 'Improving Students' Understanding of Algebra and Multiplicative Reasoning: Did the ICCAMS Intervention Work?', in Pope (ed.) *Proceedings of the 8th British Congress of Mathematics Education*. www.bsrlm.org.uk

Hodgen, Foster, Marks and Brown (2018) 'Evidence for Review of Mathematics Teaching: Improving Mathematics in Key Stages Two and Three', Evidence Review, London: Education Endowment Foundation. https://educationendowmentfoundation.org.uk/evidence-summaries/evidence-reviews/improving-mathematics-in-key-stages-two-and-three/

Holmqvist (2017) 'Models for Collaborative Professional Development for Teachers in Mathematics', *International Journal for Lesson and Learning Studies*, 6, pp. 190–201.

Howard and Miller (2018) 'Pay-for-Performance Reform Programs: It's More Than the Money!' *Urban Education*.

Hume and Coll (2009) 'Assessment Of Learning, For Learning, and As Learning: New Zealand Case Studies', *Assessment in Education: Principles, Policy and Practice*, 16, pp. 269–290.

Humphrey, Lendrum, Ashworth, Frearson, Buck and Kerr (2016) 'Implementation and Process Evaluation (IPE) for Interventions in Educational Settings: an Introductory Handbook', London, Education Endowment Foundation.

Ikemoto, Steele and Pane (2016) 'Poor Implementation of Learner-Centered Practices: A Cautionary Tale', *Teachers College Record*, 118, pp. 1–34.

Kale and Selmer (2014) 'Guiding the Development of Practice-Oriented Teacher Knowledge', *International Journal of Adult, Community and Professional Learning*, 20, pp. 25–37.

Keast (2015) 'An Effective Model for Professional Development: A Case Study of a Mathematics Teacher's Change in Australia', *International Journal of Humanities Education*, 12, pp. 1–11.

Kiemer, Gröschner, Pehmer and Seidel (2015) 'Effects of a Classroom Discourse Intervention on Teachers' Practice and Students' Motivation to Learn Mathematics and Science', *Learning and Instruction*, 35, pp. 94–103.

Kilde and Gonzales (2015) 'A Connective MOOC for K-12 Science and Mathematics Teacher Professional Development in Native American Pueblo Schools', ICTD '15: Proceedings of the Seventh International Conference on Information and Communication Technologies and Development (May 2015).

Kraft, Blazar and Hogan (2018) 'The Effect of Teaching Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence', *Review of Educational Research*, 88.

Kramarski (2009) 'Developing a Pedagogical Problem Solving View for Mathematics Teachers with Two Reflection Programs', *International Electronic Journal of Elementary Education*, 2, pp. 137–153.

Kramarski and Revach (2009) 'The Challenge of Self-Regulated Learning in Mathematics Teachers' Professional Training', *Educational Studies in Mathematics*, 72, pp. 379–399.

Lau and Yuen (2013) 'Learning Study in Mathematics: It is for Students, Teachers, and Teacher Educators', *Asia-Pacific Education Researcher*, 22, pp. 377–388.

Lee (2014) 'Conceptual Framework of Blended Professional Development for Mathematics Teachers', *Journal of Asynchronous Learning Network*, 17, pp. 81–92.

Li, Klahr and Siler (2006) 'What Lies Beneath the Science Achievement Gap: The Challenges of Aligning Science Instruction with Standards and Tests', *Science Educator*, 15, pp. 1–12.

Li and Qi (2011) 'Online Study Collaboration to Improve Teachers' Expertise in Instructional Design in Mathematics', *ZDM - International Journal on Mathematics Education*, 43, pp. 833–845.

Linacre (2000) 'Comparing "Partial Credit" and "Rating Scale" Models', *Rasch Measurement Transactions*, 14, p. 768.

Linacre (2011) 'Winsteps® Rasch Measurement Computer Program (Version 3.72.3)', Beaverton, Oregon: Winsteps.com.

Lomibao (2016) 'Enhancing Mathematics Teachers' Quality Through Lesson Study', *SpringerPlus*, 5.

Marshall and Drummond (2006) 'How Teachers Engage with Assessment for Learning: Lessons from the Classroom', *Research Papers in Education*, 21, pp. 133–149.

Marton and Booth (1997) *Learning and Awareness*, New Jersey: Lawerence Erlbaum Associates.

Marton and Säljö (1976) 'On Qualitative Differences in Learning: I – Outcome and Process', *British Journal of Educational Psychology*, 46, pp. 4–11.

McNeill, Butt and Armstrong (2016) 'Developing Collaborative Approaches to Enhance the Professional Development of Primary Mathematics Teachers', *Education 3–13*, 44, pp. 426–441.

Meletiou-Mavrotheris, Mavrou, Stylianou, Mavromoustakos and Christou (2014) 'Teaching Mathematics with Tablet PCs: A Professional Development Program Targeting Primary School Teachers', STEM Education: Concepts, Methodologies, Tools, and Applications.

Messick (1989) 'Validity', in Linn (ed.), *Educational Measurement* (3rd edn), USA: American Council on Education and Oryx Press.

Meyers, Molefe, Brandt, Zhu and Dhillon (2016) 'Impact Results of the Emints Professional Development Validation Study', *Educational Evaluation and Policy Analysis*, 38, pp. 455–476.

Mokhele and Jita (2012) 'When Professional Development Works: South African Teachers' Perspectives', *Anthropologist*, 14, pp. 575–585.

Muijs, Kyriakides, van der Werf, Creemers, Timperley and Earl (2014) 'State of the Art: Teacher Effectiveness and Professional Learning', *School Effectiveness and School Improvement*, 25 (2), pp. 231–256.

Muijs and Reynolds (2011)'*Effective Teaching: Evidence and Practice*, London, Sage.

NCETM (2009) 'Final Report: Researching Effective CPD in Mathematics Education (RECME)', London: National Centre for Excellence in Teaching Mathematics.

Ostermeier, Prenzel and Duit (2010) 'Improving Science and Mathematics Instruction: The Sinus Project as an Example for Reform as Teacher Professional Development', *International Journal of Science Education*, 32, pp. 303–327.

Pampaka, Pepin and Sikko (2015) 'Supporting or Alienating Students During Their Transition to Higher Education: Mathematically Relevant Trajectories in the Contexts of England and Norway', *International Journal of Educational Research*, 79, pp. 240–257.

Pampaka and Williams (2016) 'Mathematics Teachers' and Students' Perceptions of Transmissionist Teaching and Its Association with Students' Dispositions', *Teaching Mathematics and Its Applications: An International Journal of the IMA*, 35, pp. 118–130.

Pampaka, Williams and Homer (2016a) 'Is the Educational "What Works" Agenda Working? Critical Methodological Developments', *International Journal of Research and Method in Education*, 39, pp. 231–236.

Pampaka, Williams and Homer (2016b" 'Is the Educational "What Works" Agenda Working? Critical Methodological Developments', *International Journal of Research and Method in Education*, 39, pp. 345–348.

Pampaka, Williams and Hutcheson (2012a) 'Measuring Students' Transition Into University and Its Association with Learning Outcomes', *British Educational Research Journal*, 38, pp. 1041–1071.

Pampaka, Williams, Hutcheson, Black, Davis , Hernandez-Martinez and Wake (2013) 'Measuring Alternative Learning Outcomes: Dispositions to Study in Higher Education', *Journal of Applied Measurement,* 14, pp. 197–218.

Pampaka, Williams, Hutcheson, Wake, Black, Davis and Hernandez-Martinez (2012b) 'The Association Between Mathematics Pedagogy and Learners' Dispositions for University Study', *British Educational Research Journal*, 38, pp. 473–496.

Pampaka and Wo (2014) 'Revisiting Mathematical Attitudes of Students in Secondary Education', in Liljedahl, P., Oesterle, S., Nicol, C., and Allan, D. (eds) *Proceedings of the Joint Meeting of PME 38 and PME-NA 36*, (vol. 4, pp. 385–392, in Liljedahl, Oesterle, Nicol and Allan (eds), *The Joint Meeting of PME 38 and PME-NA 36*, Canada, Vancouver. PME, pp. 385–392.

Panizzon and Pegg (2008) 'Assessment Practices: Empowering Mathematics and Science Teachers in Rural Secondary Schools to Enhance Student Learning', *International Journal of Science and Mathematics Education*, 6, pp. 417–436.

Parada and Pluvinage (2014) 'Considerations on Didactic Thinking by Teachers of Mathematics/Reflexiones de Profesores de Matemáticas Sobre Aspectos Relacionados con su Pensamiento Didáctico', *Revista Latinoamericana de Investigación en Matemática Educativa*, 171, pp. 83–114.

Piccolo, Harbaugh, Carter, Capraro and Capraro (2008) 'Quality of Instruction: Examining Discourse in Middle School Mathematics Instruction', *Journal of Advanced Academics*, 19, pp. 376–410.

Pournara, Hodgen, Adler and Pillay (2015) 'Can Improving Teachers' Knowledge of Mathematics Lead to Gains in Learners' Attainment in Mathematics?', *South African Journal of Education*, 35.

QCA (2010) 'Assessing Pupils' Progress: Assessment at the Heart of Learning', Coventry: Qualifications and Curriculum Development Agency (QCA/08/3867): https://dera.ioe.ac.uk/10945/7/Assess_pupils_progress_webo_Redacted.pdf; http://www.qca.org.uk/7883.html

Rockoff (2004) 'The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data', *American Economic Review*, 94, pp. 247–252.

Ronfeldt, Farmer, McQueen and Grissom (2015) 'Teacher Collaboration in Instructional Teams and Student Achievement', *American Educational Research Journal*, 52, pp. 475–514.

Ruthven, Mercer, Taber, Guardia, Hofmann, Ilie, Luthman and Riga (2017) 'A Research-Informed Dialogic-Teaching Approach to Early Secondary School Mathematics and Science: The Pedagogical Design and Field Trial of the Episteme Intervention', *Research Papers in Education*, 32, pp. 18–40.

Ryan and Williams (2007) *Children's Mathematics 4–15*, Milton Keynes: Open University Press.

Schoenfeld (2011) 'Toward Professional Development for Teachers Grounded in a Theory of Decision Making', *ZDM - International Journal on Mathematics Education*, 43, pp. 457–469.

Shayer and Adhami (2007) 'Fostering Cognitive Development Through the Context of Mathematics: Results of the Came Project', *Educational Studies in Mathematics*, 64, pp. 265–291.

Shulman (1986) 'Those Who Understand: Knowledge Growth in Teaching', *Educational Researcher*, 15, pp. 4–14.

Shulman (1987) 'Knowledge and Teaching: Foundations of the New Reform', *Harvard Educational Review*, 57, pp. 1–22.

Skott (2019) 'Changing Experiences of Being, Becoming, and Belonging: Teachers' Professional Identity Revisited', *ZDM - Mathematics Education*, 51, pp. 469–480.

Slade (2009) 'Assessing Pupils' Progress in Science', *Education in Science*, 231, pp. 10–11.

Slavin, Lake and Groff (2009) 'Effective Programs in Middle and High School Mathematics: A Best Evidence Synthesis', *Review of Educational Research*, 79, pp. 839 - 911.

Smith and Gorard (2005) '"They Don't Give Us Our Marks": The Role of Formative Feedback in Student Progress', *Assessment in Education: Principles, Policy and Practice*, 12, pp. 21–38.

Spybrook and Raudenbush (2009) 'An Examination of the Precision and Technical Accuracy of the First Wave of Group-Randomized Trials Funded by the Institute of Education Sciences', *Educational Evaluation and Policy Analysis*, 31, pp. 298–318.

Stein, Engle, Smith and Hughes (2008) 'Orchestrating Productive Mathematical Discussions: Five Practices for Helping Teachers Move Beyond Show and Tell', *Mathematical Thinking and Learning*, 10, pp. 313–340.

Stenhouse (1975) *An Introduction to Curriculum Research and Development*, London, Heinemann.

Streefland (1991) *Fractions in Realistic Mathematics Education: A Paradigm of Developmental Research*, Dordrecht, Kluwer Academic.

Swan (2006) 'Designing and Using Research Instruments to Describe the Beliefs and Practices of Mathematics Teachers', *Research in Education*, pp. 58–70.

Takahashi (2011) 'The Japanese Approach to Developing Expertise in Using the Textbook to Teach Mathematics', *Expertise in Mathematics Instruction: an International Perspective*.

Takahashi and McDougal (2016) 'Collaborative Lesson Research: Maximizing the Impact of Lesson Study', *ZDM - Mathematics Education*, 48, pp. 513–526.

Tirosh, Tsamir and Levenson (2015) 'Fundamental Issues Concerning the Sustainment and Scaling Up of Professional Development Programs', *ZDM Mathematics Education*, 47, pp. 153–159.

Torgerson and Torgerson (2008) *Designing Randomised Trials in Health, Education and the Social Sciences: An Introduction*, Palgrave Macmillan.

Troncoso (2020) 'Minimum Detectable Effect Size (MDES) in a 2-Level Cluster Randomised Controlled Trial': https://github.com/patroncos/mdesapp

Tytler, Symington, Darby, Malcolm and Kirkwood (2011) 'Discourse Communities: A Framework from Which to Consider Professional Development for Rural Teachers of Science and Mathematics', *Teaching and Teacher Education*, 27, pp. 871–879.

van Galen, Feijs, Figueiredo, Gravemeijer, van Herpen and Keijzer (2008) *Fractions, Percentages, Decimals and Proportions: A Learning Teaching Trajectory for Grade 4, 5 and 6*, Rotterdam Netherlands: Sense.

van Zoest, Breyfogle and Ziebarth (2002) 'Self-Perceived and Observed Practices of Secondary School Mathematics Teachers', *Teacher Development*, 6, pp. 245–268.

Warwick, Vrikki, Færøyvik Karlsen, Dudley and Vermunt (2019) 'The Role of Pupil Voice as a Trigger for Teacher Learning in Lesson Study Professional Groups', *Cambridge Journal of Education*, 49, pp. 435–455.

Warwick, Vrikki, Vermunt, Mercer and van Halem (2016) 'Connecting Observations of Student and Teacher Learning: An Examination of Dialogic Processes in Lesson Study Discussions in Mathematics', *ZDM - Mathematics Education*, 48, pp. 555–569.

Watson (2006) 'Some Difficulties in Informal Assessment in Mathematics', *Assessment in Education*, 13, pp. 289–303.

Watson and Beswick (2011) 'School Pupil Change Associated with a Continuing Professional Development Programme for Teachers', *Journal of Education for Teaching*, 37, pp. 63–75.

Wiliam (2007a) 'Five "Key Strategies" for Effective Formative Assessment', National Council of Teachers of Mathematics (research brief).

Wiliam (2007b) 'Keeping Learning on Track: Formative Assessment and the Regulation of Learning', in Lester Jr. (ed.), *Second Handbook of Research on Mathematics Teaching and Learning*, Greenwich, CT: Information Age.

Wiliam (2011) *Embedded Formative Assessment*, Solution Tree.

Wiliam (2016) *Leadership for Teacher Learning*, West Palm Beach, FL: Learning Sciences International.

Wiliam and Thompson (2007) 'Integrating Assessment with Instruction: What Will It Take to Make It Work?, in Dwyer (ed.), *The Future of Assessment: Shaping Teaching and Learning*, Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, Black, Davis, Hernandez-Martinez, Hutcheson, Nicholson, Pampaka and Wake (2008) 'TLRP Research Briefing No 38: Keeping Open the Door to Mathematically Demanding Programmes in Further and Higher Education', School of Education, University of Manchester.

Wolfe and Smith Jr. (2007a) 'Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part I - Instrument Development Tools', *Journal of Applied Measurement*, 8, pp. 97–123.

Wolfe and Smith Jr. (2007b) 'Instrument Development Tools and Activities for Measure Validation Using Rasch Models: Part II - Validation Activities', *Journal of Applied Measurement*, 8, pp. 204–234.

Wong (2010) 'What Makes a Professional Learning Community Possible? A Case Study of a Mathematics Department in a Junior Secondary School of China', *Asia Pacific Education Review*, 11, pp. 131–139.

Wylie and Lyon (2015) 'The Fidelity of Formative Assessment Implementation: Issues of Breadth and Quality', *Assessment in Education: Principles, Policy and Practice*, 22, pp. 140–160.

Yin, Olson, Olson, Solvin and Brandon (2015) 'Comparing Two Versions of Professional Development for Teachers Using Formative Assessment in Networked Mathematics Classrooms', *Journal of Research on Technology in Education*, 47, pp. 41–70.

Zakaria and Daud (2009) 'Assessing Mathematics Teachers' Professional Development Needs', *European Journal of Social Sciences*, 8, pp. 225–231.

Appendix A: EEF cost rating

Figure 2: Cost Rating

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

Appendix B: Security classification of trial findings

OUTCOME: *MATHEMATICS ATTAINMENT (MaLT)*

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

Threats to validity	Threat to internal validity?	Comments
Threat 1: Confounding	Low	Randomisation was conducted independently using anonymised school lists. Minimal imbalance was observed at baseline across key characteristics, including pre-test KS2 score (effect size 0.045).
Threat 2: Concurrent Interventions	Moderate	Data from a subset of control schools (n=24) suggests that many schools engaged with other maths professional development approaches, including Mastery, which may have weakened experimental contrast. No information was collected from intervention schools about other programmes implemented before or alongside ICCAMS.
Threat 3: Experimental effects	Low	Business as usual data was collected from control schools at a single time point such that the authors cannot draw conclusions about potential experimental effects. School-level randomisation minimised risk of contamination between treatment and control groups.
Threat 4: Implementation fidelity	High	Overall, of 53 schools with compliance data, 8 fulfilled the highest criteria. Differences were observed between Lead and Cascade Teachers with respect to implementation, including variation in training attendance and lesson delivery. Cascade Teachers were also less likely than Lead teachers to report changes in perceptions of teaching practice in line with ICCAMS.
Threat 5: Missing Data	Moderate	The sample size at baseline is not known with certainty, but overall missing data is estimated to be 17.2%, with some imbalance across groups (19.6% treatment, 14.4% control). Multiple imputation models suggest that complete-case analysis may underestimate effect sizes, but do not change the substantive findings.
Threat 6: Measurement of Outcomes	Low	The primary outcome uses a slightly modified version of the Mathematics Assessment for Learning and Teaching (MaLT) for Year 8,

		measured in terms of both raw and Rasch scores. Revisions to the assessment were piloted and independently reviewed. Tests were conducted under exam conditions with administrators and markers blind to condition. No floor/ceiling effects observed.
Threat 7: Selective reporting	Low	The trial is registered and pre-specified. Primary analysis follows the published statistical analysis plan.

- **Initial padlock score:** [4] Padlocks – Cluster randomised trial with MDES at randomisation of 0.19 and overall attrition of 17.2%.
- **Reason for adjustment for threats to validity**: [-1] Padlocks – Three threats to validity identified as moderate to high and all suggest that the impact estimate may be underestimated. Reduction of one padlock.
- **Final padlock score:** Initial score adjusted for threats to validity = [3] Padlocks

Appendix C: Effect size estimation

Table D1: Effect size calculations for primary and secondary outcomes

					Conditional Full Model		Null Model (Empty)	
Outcome	Condition Coefficient	CI-Low	CI-UP	p-value	Variance-School	Variance-student	Variance-School	Variance-student
Total Math Score	0.232	-0.454	0.918	0.507	2.884	36.318	8.523	103.832
Algebra-Raw	0.119	-0.097	0.335	0.280	0.282	4.135	0.703	8.375
Multiplication-Raw	0.095	-0.178	0.367	0.496	0.438	8.390	1.475	21.957
Maths Disposition	0.048	-0.037	0.134	0.270	0.038	1.253	0.068	1.661

Effect size calculations - ITT

					Conditional Full Model		Null Model (Empty)	
Outcome	Condition Coefficient	CI-Low	CI-UP	p-value	Variance-School	Variance-student	Variance-School	Variance-student
Total Math Score	0.232	-0.454	0.918	0.507	2.884	36.318	8.523	103.832
Algebra-Raw	0.119	-0.097	0.335	0.280	0.282	4.135	0.703	8.375
Multiplication-Raw	0.095	-0.178	0.367	0.496	0.438	8.390	1.475	21.957
Maths Disposition	0.048	-0.037	0.134	0.270	0.038	1.253	0.068	1.661

Effect size calculations – Subgroup analyses

					Conditional Full Model		Null Model (Empty)	
Outcome	Condition Coefficient	CI-Low	CI-UP	p-value	Variance-School	Variance-student	Variance-School	Variance-student
Total Math Score	0.412	-0.240	1.064	0.215	1.741	41.169	3.581	94.610
Algebra-Raw	0.160	-0.067	0.387	0.167	0.282	4.135	0.222	4.524
Multiplication-Raw	0.224	-0.045	0.493	0.103	0.244	9.338	0.748	20.600
Maths Disposition	0.091	-0.031	0.213	0.143	0.050	1.403	0.109	1.689

Further appendices:

See separate document with further appendices.

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn