Maths-Whizz Intelligent Tutoring Programme

Evaluation report

June 2026

Katherine Aston, Aarti Sahasranaman,
Gemma Schwendel, and Stephen Welbourne

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

We do this by:

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

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

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

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

0207 802 1653

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Table of contents

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER). The trial director and principal investigator for this study was Stephen Welbourne, Research Director. Aarti Sahasranaman, Senior Research Manager, led the evaluation team and the impact evaluation. Katherine Aston, Research Manager, led the implementation and process evaluation (IPE). They were supported by Gemma Schwendel as trial statistician and Gustavo Lopes as IPE researcher. Kathryn Hurd led the research operations with Jishi Jose and Tom Dickinson as operations researchers.

Contact details:

National Foundation for Educational Research
The Mere,
Upton Park,
Slough,
Berkshire SL1 2DQ,
United Kingdom

Email: a.sahasranaman@nfer.ac.uk

Acknowledgements

We are grateful to all the infant, junior, and primary schools that participated in this evaluation. We are especially thankful to the school staff members who completed the baseline assessments and coordinated the endpoint assessments with NFER test administrators and the staff and pupils who took part in the case study visits and interviews.

We would like to give a special mention to the staff at Whizz Education who have worked with us to support the evaluation whenever needed, namely Sarah Hawkes, Emma Ringe, Lindsay Brownsword, Ray Douse, and Richard Marett.

We would further like to thank the team at the Education Endowment Foundation, including Sarah Conner, Andrew Kensley, Lottie Norton, and Thomas McKay.

Executive summary

The project

Maths-Whizz, developed and delivered by Whizz Education, is an Edtech programme consisting of an online intelligent tutoring system that provides personalised tutoring to pupils, alongside a library of digital resources and live reports for teachers. Teachers receive two 45-minute training sessions. The first focuses on facilitating pupil engagement with the tutor and empowering staff to understand and act upon the associated pupil engagement data. The bespoke second session focuses on using reports of pupil learning data to direct class teaching and the use of the digital resource library. Whizz Education encourage pupils to use the tutor for 45–60 minutes per week in addition to, rather than a replacement for, their usual maths teaching. In this trial, Maths-Whizz was delivered as a whole-class programme to pupils aged 6–10 years in Years 2, 3, 4, and 5 in state-maintained English schools during the 2024/2025 academic year.

Schools participating in this efficacy trial were randomly allocated to one of two arms. In Arm 1, pupils in Years 3 and 5 were allocated to receive Maths-Whizz whereas in Arm 2, pupils in Years 2 and 4 were allocated to receive Maths-Whizz. In both arms, year groups who were not allocated to receive Maths-Whizz acted as the control group for the other arm. We have coined the term 'interleaved design' to describe this arrangement. The primary outcome of the trial was pupils' maths attainment measured using the Renaissance Learning Star Maths assessment scaled score. The secondary outcome was pupils' motivation towards maths learning measured using the Maths and Me survey. The implementation and process evaluation (IPE) used surveys, interviews, and focus groups focusing particularly on exploring implementation fidelity. Whizz Education recruited 63 schools to the trial. In total, 10,369 pupils completed the baseline assessment in July 2024; 5,136 pupils were allocated to the control group, and 5,233 to the intervention group. Maths-Whizz was delivered from September 2024 – June 2025.

Table 1: Key conclusions

Key conclusions
1. Pupils allocated to receive Maths-Whizz made one additional month's progress in maths, on average, compared to pupils who were not allocated to receive the programme. This finding has high security.
2. Among children eligible for free school meals (FSM), those allocated to receive Maths-Whizz made two additional months' progress in maths, on average, compared to FSM-eligible pupils who were not allocated to receive the programme. These results may have lower security than the overall findings because of the smaller number of pupils.
3. Pupils allocated to receive Maths-Whizz reported a small positive effect on their mathematical self-efficacy and enjoyment of maths compared to pupils who were not allocated to receive the programme. This was supported by teachers' feedback that intervention pupils 'looked forward' to using Maths-Whizz. Pupils were also broadly positive about the programme and particularly enjoyed the gamified learning on Maths-Whizz.
4. On average, pupils engaged with the Maths-Whizz platform for around 32 minutes per week, with around 25 minutes spent in the tutor mode learning new content and eight minutes in the practice mode reviewing and consolidating content. Higher usage in both modes was associated with better maths attainment with time in the practice mode showing the strongest association.
5. Several features of Maths-Whizz might have contributed to this impact, particularly the larger impact for FSM-eligible pupils. These include adaptive teaching, dynamic feedback mechanisms (real time, adaptive responses that continuously adjust questions, difficulty, and support based on a student's performance in order to personalise learning), self-paced individualised learning experiences, and making gamified learning accessible, all of which make learning more accessible and equitable for disadvantaged pupils. The stronger impact for FSM-eligible pupils suggests that this platform may be particularly adept at identifying and addressing gaps in learning, which are more prevalent among FSM-eligible children.

EEF security rating

These findings have a high security rating. This was an efficacy trial, which tested whether the intervention worked under developer-led conditions in a number of schools. The trial was a well-designed two-armed interleaved randomised controlled trial. The trial was well powered. The pupils in Maths-Whizz year groups were similar to those in the comparison groups. However, 13.4% of the pupils who started the trial were not included in the final analysis. The reasons for this attrition included the withdrawal of one school from the study, pupil absence on test day, pupils leaving the school, withdrawal of pupils from testing by the teacher because of additional needs, and internet connectivity issues.

Additional findings

Pupils who were allocated to receive Maths-Whizz across Years 2, 3, 4, and 5 made, on average, one additional month's progress than those who did not receive the programme. This is our best estimate of impact, which has a high security rating. As with any study, there is always some uncertainty around the result: the possible impact of Maths-Whizz also included no progress and up to two additional months' progress. Pupils eligible for FSM who received Maths-Whizz made, on average, two additional months' progress compared with FSM-eligible pupils who were not allocated to receive the programme. This suggests that while all children benefited from Maths-Whizz, FSM pupils benefited more than non-FSM pupils. As always, there is some uncertainty around this result: the possible impact of Maths-Whizz for FSM-eligible pupils also includes a positive impact of one to three additional month's progress. Maths-Whizz also had a positive impact on pupils in each year group, with the largest impact on Year 3 pupils who made the equivalent of two additional months' progress.

Around 93% of surveyed teachers reported being satisfied with the Maths-Whizz programme. Most teachers perceived that Maths-Whizz had impacted their pupils positively (e.g. reducing gaps in learning, increasing pupils' willingness to give it a go). Pupils also reported enjoying using Maths-Whizz, particularly the gamified learning aspects of the programme. The issues pupils flagged were related mostly to technology (e.g. connectivity, device access issues) and to a smaller extent to not being able to achieve progressions due to the duration of the session.

Cost

The average cost of delivering Maths-Whizz for two of the year groups in a school was around £1,785, or £21.51 per pupil per year when averaged over three years.

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
Maths attainment (all pupils)	0.08 (0.02, 0.14)	1	4 of 5 padlocks	8,983	0.01	£ £ £ £ £ (1 of 5)
Maths attainment (FSM pupils)	0.14 (0.06, 0.22)	2	4 of 5 padlocks	2,301	0.01	£ £ £ £ £ (1 of 5)

Introduction

Background

According to the Department for Education's (DfE's) latest publication on attainment for the Key Stage 2 National Curriculum assessments in England, in the 2024/2025 academic year, 61% of pupils from disadvantaged backgrounds met the expected standard in maths compared to 80% of pupils from non-disadvantaged backgrounds (DfE, 2025). The COVID-19 pandemic, while damaging to the maths learning of all pupils (Renaissance Learning and Education Policy Institute, 2021; EEF, 2022), has only further widened the attainment gap between disadvantaged pupils and their more affluent peers. A study commissioned by the Education Endowment Foundation (EEF) estimated that between the Autumn Term of 2019 and the end of the 2020/2021 academic year, the maths attainment gap between disadvantaged pupils and their peers had widened by an equivalent of one additional month's progress (EEF, 2022a). Maths attainment at primary school is crucial because the links between Key Stage 2 maths performance and later educational achievement are particularly strong (Menzies, Ramaiah and Boulton, 2021). This means that pupils who perform poorly in maths at primary school are unlikely to be able to turn this around and perform better at the end of secondary school, with unfavourable downstream social and economic consequences (YouGov, 2014; Hodge, Little, and Weldon, 2021).

Tutoring, when targeted to disadvantaged pupils, could provide an effective means to increase attainment and close the attainment gap between disadvantaged pupils and their peers. In fact, tutoring was widely adopted as an academic catch-up intervention in response to the COVID-19 pandemic. The DfE, for example, launched the National Tutoring Programme, with the stated aim of providing high-quality tutoring to disadvantaged pupils in England (DfE, 2022). Evidence shows that one to one tutoring is effective at improving pupil outcomes with pupils making, on average, an additional five months' progress (Harrison and Higgins, 2023). These improvements, however, vary substantially depending on the subject, with the effects in maths appearing to be lower (additional two months' progress) compared to literacy (additional six months' progress). Delivering one to one tuition is expensive, and this is especially true when tutoring is delivered by qualified teachers. Small group tuition, in comparison, has a slightly smaller impact, with an average of an additional four months' progress (Harrison and Higgins, 2023) but may be a cost-effective solution for schools. As with one to one tutoring, the impact of small group tuition varies depending on the subject. Most of the research on small group tuition has focused on reading, where the effect is higher (additional four months' progress) compared to maths (additional three months' progress). However, the increasing demand for tutoring has not been commensurate with the supply of qualified tutors, particularly in maths (Worth and McLean, 2025).

Artificial intelligence (AI) enabled tutoring systems show promise both in terms of offering a more affordable tutoring option and sidestepping the requirement for a qualified human tutor. They also have the potential to support teachers who face the challenging and time-consuming task of teaching pupils with widely varying prior attainment within the same classroom. For instance, research by Whizz Education (developers of the Maths-Whizz programme) based on a representative sample of 400 UK primary classrooms using Maths-Whizz has shown that by upper primary there is a four-year knowledge gap between learners in a classroom (Whizz Education, 2021). AI-enabled adaptive tutoring systems can simulate human tutors by providing personalised tutoring, adapting to the individual needs of the pupil and providing instant feedback. A systematic review of maths apps for children aged four to seven found that personalised, adaptive features are more effective at maximising children's maths attainment than non-adaptive approaches (Outhwaite *et al.,* 2022), highlighting the promise of intelligent tutoring systems. However, while there is a plethora of evidence demonstrating the impact of human led tutoring, it is unknown if this would translate to AI enabled tutoring systems – this is the first UK RCT study investigating this at any reasonable scale. It should also be noted that even within human based tutoring the actual impact is likely to be heavily influenced by implementation factors such as dosage and whether the tutoring supplements or replaces class teaching. These factors would be equally relevant to AI enabled tutoring so any comparison between human and AI tutoring would need to carefully control for these.

Maths-Whizz, developed and delivered by Whizz Education,[1] is an Edtech programme consisting of an online intelligent tutoring system, a library of digital resources for teachers and live reports that support decision-making by teachers. Schools receive implementation support from Whizz Education, which includes training on the use of the platform, regular monitoring of platform usage, and termly and annual progress reviews. Maths-Whizz is currently active in 21 countries and has reached over 9,000 schools and tutored 1.5 million pupils.

A quasi-experimental study of the impact of Maths-Whizz on maths attainment of pupils showed an effect size equivalent to an additional three months' progress (Mavrikis, Schlepps and Mubeen, 2018). The study involved 3,400 pupils in Mexico aged eight to nine years (fourth grade) who had access to the Maths-Whizz platform for seven weeks. This efficacy trial was the first to evaluate the impact of the Maths-Whizz Intelligent Tutoring programme in English schools using a robust experimental design. This randomised controlled trial (RCT) was larger in scope than the previous study and involved pupils in Years 2, 3, 4, and 5 (ages six to ten) in 64 primary schools in England who received access to the Maths-Whizz programme for the whole 2024/2025 academic year. Focusing on multiple year groups allowed us to adopt an interleaved design, which maximises statistical power and is more attractive to schools as all schools receive the intervention in some year groups. It also allowed us to evaluate the effect of the intervention in a range of age groups. Schools were randomly assigned to either receive the intervention in Years 2 and 4 OR in Years 3 and 5. The year groups not receiving the intervention in each arm served as the control group for the year groups receiving the intervention in the other arm. The primary outcome was maths attainment measured using the Renaissance Learning Star Maths assessment (henceforth referred to as the Star Maths assessment). The secondary outcome was pupil motivation towards maths learning measured using the Maths and Me survey. The integrated implementation and process evaluation (IPE) explored how, and in what circumstances, Maths-Whizz impacts pupils and teachers. We explored these questions using a combination of qualitative and quantitative methods. The qualitative methods consisted of interviews with teachers and senior leaders and focus groups with pupils while the quantitative methods involved analyses of surveys of teachers and of Maths-Whizz platform data. The cost evaluation provides a robust estimate of the cost to schools per pupil per year of delivering the Maths-Whizz programme.

Intervention

Maths-Whizz is an adaptive virtual tutor developed by Whizz Education. This AI-enabled tutor is designed for children between ages five and 13 years. A detailed description of the intervention in the context of the Template for Intervention Description and Replication (TIDieR) checklist is presented in Table 3 below.

Table 3: TIDieR checklist for Maths-Whizz

TIDieR item	Description
Brief name	Maths-Whizz Intelligent Tutoring Programme
Why	One to one tutoring has been shown to be effective in improving pupil outcomes (Harrison and Higgins, 2023), but is comparatively expensive and depends on a supply of qualified tutors. AI-enabled tutoring platforms such as Maths-Whizz simulate human tutors by providing personalised tutoring, adapting to the individual needs of the pupil and providing instant feedback. A systematic review of maths apps for children aged four to seven found that personalised, adaptive features are more effective at maximising children's maths attainment than non-adaptive approaches (Outhwaite *et al.,* 2022), highlighting the promise of intelligent tutoring systems. A quasi-experimental study, which evaluated the impact on maths attainment of seven weeks' access to Maths-Whizz for fourth-grade pupils in Mexico, showed an effect size equivalent to an additional three months' progress (Mavrikis, Schlepps and Mubeen, 2018).
What (materials)	All schools must purchase a School Solution, which provides access to the Teachers' Resource, Teacher Platform, for an unlimited number of teachers within the school, along with training. Schools in the trial used the enhanced support model, which provides extensive ongoing implementation support throughout the year. In addition to the School Solution, each school must also purchase individual subscriptions for each pupil who will access the Maths-Whizz platform. Parent access is provided via the pupil access so one parent gains access to the parent platform for each individual pupil subscription activated. All schools participating in the trial paid a subsidised fee of £500 to access both the School

[1] See: www.whizzeducation.com/

Solution and the Maths-Whizz programme for all participating pupils in two of the year groups (i.e. either Years 2 and 4 OR Years 3 and 5).

All participating pupils are provided with a login and password to access their personal Maths-Whizz accounts and the content within. This includes lessons for each learning objective, practice exercises, tests, and Topic Challenges to review and consolidate their learning.

Each pupil's journey on Maths-Whizz begins with an interactive and fully adaptive initial assessment to diagnose the gaps in their knowledge and determine their current proficiency across several key topics in the maths curriculum. The tutor then creates a learning and support plan for each pupil tailored to their specific needs based on their 'maths age'. 'Maths age' is defined by Whizz Education as a measure of a 'child's maths ability against the expected level of an average student of their age'. In addition to the overall 'maths age', the tutor will also calculate pupils' 'maths age' for each topic (called Topic Age) in the maths curriculum.

Maths-Whizz provides personalised tutoring to pupils in Reception to Year 8 and covers topics ranging from place value and properties of numbers to shape and space, and probability.

The topics for Reception to Year 6 are fully aligned to the national curriculum learning objectives. Maths-Whizz is aligned to 96% of the learning objectives up to Year 8.

The Maths-Whizz tutor is underpinned by the Maths-Whizz curriculum, which comprises 1,200+ individual learning objectives covering the 5–13 age range. Learning objectives are derived from previous and current curricula frameworks, as well as Whizz's general understanding of mathematical concepts. To enable a curriculum structure, Whizz groups each lesson into one of 22 topics. Within a topic, lessons are strictly ordered. The exact position of a given lesson in its topic depends on the age at which pupils are expected to encounter that learning objective.

Each Whizz learning objective is covered by a Whizz lesson. A Whizz lesson consists of three parts:

- **A Teaching Page.** Introduces the underlying concept to pupils.
- **An interactive exercise.** This is usually made up of ten questions along with help for when pupils get stuck.
- **A test.** Mimics a worksheet and usually consists of five questions.

Teachers receive access to their own accounts where they will be able to view and monitor the progress of all pupils in their class. Teachers also receive access to a digital library, Maths-Whizz Teachers' Resource containing over 1,200 instructional and assessment resources to supplement the Maths-Whizz tutor. These resources are designed to be used with the whole class or with individual pupils and are aligned to the national curriculum. The types of resources provided include:

- **Lessons.** Engaging, animated, and interactive exercises and tests, to use in class, organised by Topics or year group to support lesson planning.
- **Activities.** A wide range of engaging problems, challenges, activities, and games to complement learning in the maths lesson.
- **Worksheets.** Includes printable worksheets to consolidate learning and support formative and summative assessments.
- **Enrichment sheets.** Rich tasks to develop mathematical reasoning, and problem-solving skills to develop fluency, application, and to challenge pupils.
- **Classroom resources.** A range of key resource and learning tools to support and enhance the learning of maths concepts.

In addition, using the comprehensive live learning data (automatically generated as pupils engage with Maths-Whizz) and the associated reporting platform, teachers can identify learning gaps and provide further targeted intervention where needed. Live learning data can also inform decision-making for teaching and learning improvements, curriculum delivery, and education strategy.

Parents receive access to a parent account where they can engage with their child's learning and monitor their child's progress and needs.

What (procedures)	Following recruitment, Whizz Education develops a bespoke implementation plan aligned to each school's goals and to gain leadership support to ensure implementation is ingrained within the school's core maths provision. To initiate this implementation planning process, Whizz Education sets up a 45–60-minute online session with the school leadership and maths lead to determine the school's vision for success, long-term and short-term goals, agree training dates and action milestones, and reinforce high-level expectations. All schools receive training and onboarding from a Whizz Education Success Partner (ESP) to enable school leadership and teaching staff to facilitate pupil engagement with the tutor and empower staff to understand and act upon the associated learning analytics. • **Session 1.** This 45-minute online session includes expectation setting, basic knowledge of the Maths-Whizz platform (e.g. completing assessments, adaptive nature of the tool, what a typical lesson looks like and an overview of the teacher admin platform). It is conducted as soon as possible upon starting (for this efficacy trial, this was in August/September 2024) with teachers of all participating classes and at least one member of the leadership team. • **Session 2.** This online session is tailored around each school's needs. There is a focus on the teacher platform and teachers' data feed, using the Maths-Whizz platform to improve quality first teaching, a deeper dive into class reports, interpreting and actioning data, and how to use the tools such as Teachers' Resources and Topic Focus to support teaching. This 45-minute session with the maths lead (and teachers of participating classes subject to availability) is conducted six to eight weeks following the completion of initial assessments. The ESP provides ongoing support to schools throughout implementation of the programme.
Who provided (intervention providers/implementers)	The basic concepts and design of Maths-Whizz were originally developed by graphic designer and author of interactive maths books, Ron van der Meer, and co-author and teacher Bob Gardner. Since then, it has continued to be adapted and enhanced by an in-house research and product development team at Whizz Education including science, technology, engineering, and mathematics (STEM) graduates from Oxford, Harvard, and Imperial College London. Whizz Education provides access to schools, teachers, pupils, and parents to their respective areas of the Maths-Whizz platform. For each class, teachers can either allow the tutor to determine the topics to be covered or they can choose Topic Focus to direct Maths-Whizz to directly support the topic they are teaching in class. When teachers choose Topic Focus, Maths-Whizz will set-up differentiated lessons for each pupil in the class. Teachers can access assessments and progress reports to change course for each pupil depending on their changing needs and learning progress. Whizz Education also accesses the pupil, class, and school assessments and progress reports to provide schools with targeted support and advice to optimise implementation.
Who (recipients)	The Maths-Whizz virtual intelligent tutor provides personalised tutoring to pupils from Reception to Year 8. The tutor can be implemented to the whole class or to a subset of pupils, but access to the platform is on an individual basis. For the purpose of this efficacy trial, Maths-Whizz was available to the whole participating class of pupils in Years 2, 3, 4, and 5. While there are no specific exclusion criteria, pupils with severe visual disabilities were unable to engage with the tutor. Whizz Education, the delivery partner, recruited schools to participate in the trial. The eligibility criteria for schools were: • to be an infant/junior/primary school with children in Years 2, 3, 4, and 5 as of 01 September 2024; • not to have mixed year maths teaching even if the school has mixed year groups (i.e. children who are from more than one year group within the same class) in any of Years 2, 3, 4, and 5 in the 2024/2025 academic year; • not to have implemented Maths-Whizz or any Whizz associated services in the 2023/2024 academic year; • not be participating in another funded trial by the EEF in the 2024/2025 academic year;

	• to have a suitable level of IT provision including devices such as tablets, laptops, desktops, and sufficient Wi-Fi™ capacity (0.5 Mbps internet connection per concurrent pupil) (see Maths-Whizz minimum requirements); and • to agree to contribute £500 towards the cost of the programme.
How (mode of delivery), when and how much	For this efficacy trial, pupils were provided access to the Maths-Whizz intelligent tutor throughout the 2024/2025 academic year, i.e. starting in September 2024 till July 2025. Schools had the opportunity to discuss with Whizz Education the option of providing access to Maths-Whizz to pupils beyond the trial period. Pupils engage in individualised learning through the Maths-Whizz online intelligent tutor. Lessons are automated and differentiated for every individual pupil depending on their unique needs and pace of learning. • When a pupil first logs in they complete a low pressure, interactive, and adaptive initial assessment to identify individual strengths, weaknesses, and gaps in knowledge. • Maths-Whizz then provides interactive lessons, exercises, and tests in a sequence appropriate to each individual pupil's proficiency, constantly adapting at their own pace of learning with a focus on plugging knowledge gaps to accelerate pupil learning outcomes, reduce maths anxiety, and build confidence. • All learning objectives are presented in three parts: i) an animated lesson/tutorial that introduces the new concept or method; ii) an interactive exercise to determine whether the learner has internalised the concept by answering a set of questions and receiving help as needed; and iii) a brief test to establish the learner's level in the topic. Pupils only progress once they have passed the test, reflecting a mastery approach to progression. • As pupils complete learning objectives, their progress is available for teachers to monitor in a central reporting platform. The reporting platform empowers teachers, providing actionable information to support lesson planning for teachers, and further pupil intervention, including a wide range of usage and progress metrics. The programme consists of engaging pupils with the Maths-Whizz tutor for about an hour each week. This can be divided into several sessions, such as three times a week (e.g. 3 x 20-minute sessions per week), and is provided in addition to regular maths lessons. In addition to being used in school time, Maths-Whizz may also be deployed within the school's homework policy. Details of how each school will implement Maths-Whizz, including the intended approach to home use were determined during the implementation planning stage. Whizz Education shared regular reports of platform usage data for each pupil in the intervention year group with the National Foundation for Educational Research (NFER). This was used to calculate dosage of the intervention received by each pupil and its impact on their maths outcomes.
Where (Types of locations/necessary infrastructure)	The trial was open to primary schools across all regions of England. Schools were required to have sufficient number of devices such as laptops and/or tablets and sufficient Wi-Fi™ capacity. The minimum IT requirements for Maths-Whizz are detailed below: • Chrome 70+/MS Edge 90+/Safari 12+/Firefox 78+/Samsung Internet 11+. • Minimum 7.9" screen (recommended for usability reasons). • WebGL-enabled graphics card and browser. • 2GB RAM. • 0.5 Mbps internet connection per concurrent pupil.
Tailoring (Adaptation of the intervention)	The implementation of Maths-Whizz is tailored to each school's needs and goals during the implementation planning process. The Maths-Whizz intelligent tutor adapts to the individual needs and pace of learning of each pupil. The tutor maps out a tailored learning journey for each pupil based on the gaps identified in the initial assessment, which is interactive and adaptive. • Maths-Whizz calculates an overall 'maths age' for each pupil, defined by Whizz Education as a measure of a 'child's maths ability against the expected level of an average student of their age'.

	• Teachers can access assessments and reports to identify further gaps and more targeted intervention where required. • Even when teachers choose to use Topic Focus to direct Maths-Whizz to align with topics being taught in class, Maths-Whizz will set-up differentiated lessons for each pupil. • The programme can be offered to pupils in school or as part of their home learning, in which case pupils will access the Maths-Whizz platform at home.
Strategies to support implementation	• Whizz Education develops a bespoke implementation plan with the school leadership and maths lead of each school that takes into account each school's needs and goals. • ESPs provide continued implementation support to schools. ESPs monitor key school metrics on a weekly basis to identify schools that require additional support. • Schools nominate a lead staff member to support the effective implementation of the programme in schools and communicate with the Whizz ESP. • Teachers are able to access the initial assessment and learning data of all pupils to determine each child's needs and progress. • Parents are also able to engage with their child's learning by monitoring their needs and progress. • ESPs organise termly check-ins with all schools to review and adjust their implementation plans, if required. • Whizz Education organise an annual review meeting to review progress against the implementation plan and share key insights.

Logic model

The logic model for Maths-Whizz is shown in Figure 1. It outlines the target population of the intervention and the activities, outputs, short-term and intermediate outcomes that lead to the outcomes both immediately after the intervention, and longer term. The primary causal pathway for the trial (shown by the yellow arrow in Figure 1) is that pupils' use of the Maths-Whizz tutoring platform for 45–60 minutes per week throughout one academic year is expected to result in improved attainment, enjoyment, and self-efficacy in maths, and improved engagement with maths teaching as long-term outcomes for pupils. Attainment, enjoyment, and self-efficacy were measured by the impact evaluation, whereas engagement with maths teaching was covered in the IPE. The Maths-Whizz tutoring platform embeds several pedagogical approaches (individualised curriculum based on learning gaps, individualised scaffolding and feedback, mastery approach, and engaging design, activities, and gamification), which are expected to lead to improved pupil outcomes. As pupil responses to these pedagogical approaches had not yet been tested in the context of Maths-Whizz, they were a key area to explore through the IPE.

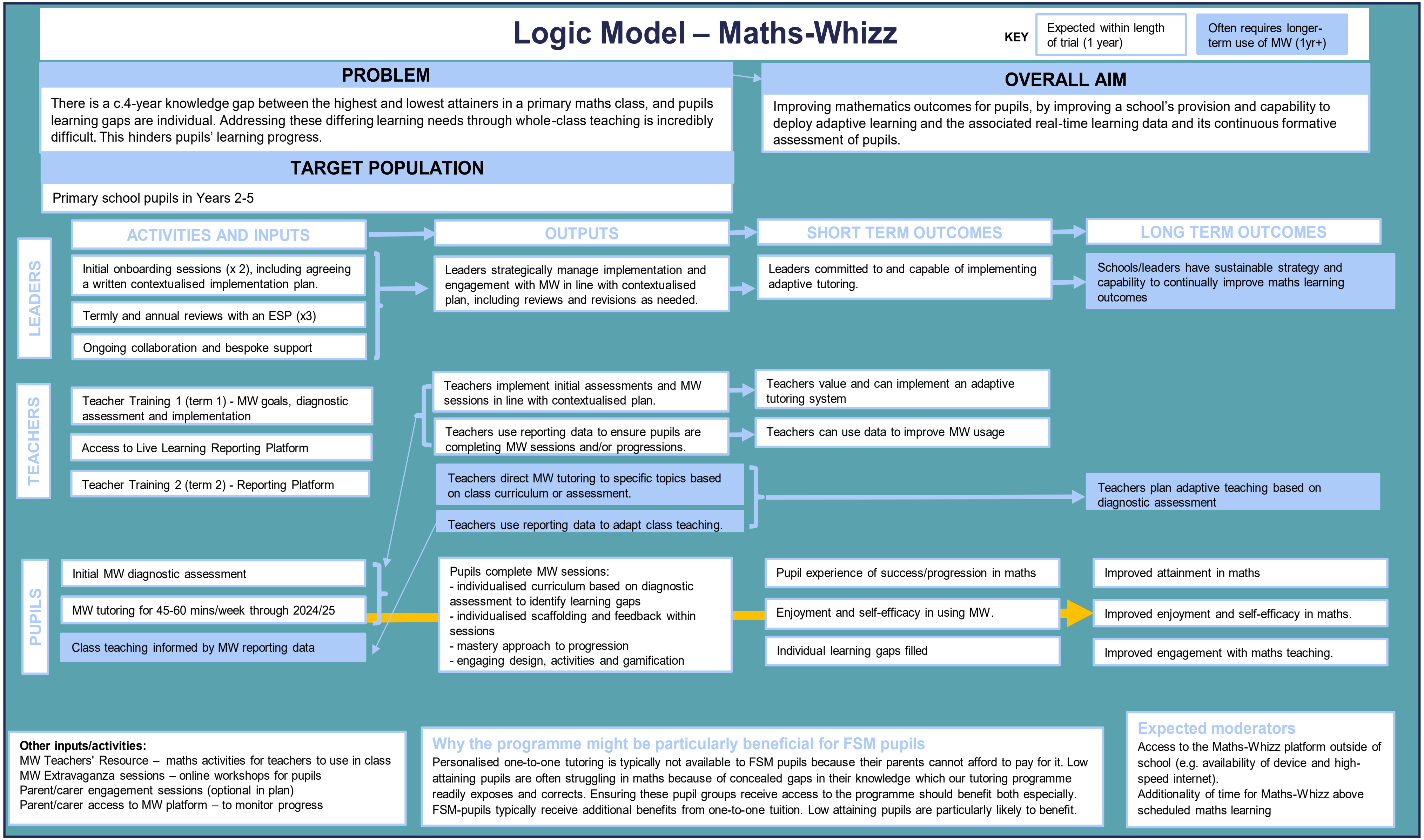
The expectation of the programme was that Maths-Whizz should be used in addition to timetabled maths lessons, rather than a replacement for these lessons. However, it was anticipated that Maths-Whizz might replace some maths learning time outside timetabled lessons, for example, maths homework, pre-teaching or pre-/post-school learning. We expected pupil impact to be moderated by whether Maths-Whizz use constitutes *additional* maths learning time or *replaces* other forms of maths learning time and this was assessed in the IPE. We also expected that where a school's implementation model includes pupil use of Maths-Whizz at home, impact would be moderated at pupil level by digital access at home (e.g. availability of a suitable device, and high-speed internet). We collected preliminary data within the IPE evaluation to explore the prevalence of this moderator, by collecting teachers' perceptions of barriers for pupils in using Maths-Whizz at home, and by monitoring any differences in usage and progression for: i) schools implementing home use; and ii) free school meals (FSM) pupils compared with non-FSM pupils, given likely differences in digital access.

Maths-Whizz provides inputs for school leaders and teachers (Activities and Inputs for Leaders and Teachers), which are captured separately in the logic model. As sustained pupil use of Maths-Whizz is the primary hypothesised pathway to impact, initial leader/teacher inputs are designed to support practical implementation, and develop buy-in/engagement from staff, to facilitate sustained pupil use. The support outlined in the logic model assumes relative stability of staffing. However, responsive support and termly reviews may compensate to some extent for changes in staffing. Within the delivery model being tested in the efficacy trial, Whizz provided extensive support for leaders and teachers, so it was important to understand what forms of support work, for which school staff, in what contexts. The IPE explored the perceived impact of different aspects of support. It also explored whether high levels of engagement with Maths-Whizz support was associated with sustained use of Maths-Whizz, to inform the prioritisation of school support for potential scale-up.

Core use of Maths-Whizz (white boxes in Figure 1) consists of implementing and monitoring sustained pupil use of the Maths-Whizz programme, which is expected to improve pupil outcomes, and to facilitate staff commitment to adaptive tutoring. Whizz expected that the trial period (one academic year) would be sufficient for schools to implement core use and achieve those outcomes.

Some uses and outcomes of Maths-Whizz are typically associated with longer time frames than the one-year trial period (blue boxes in Figure 1). Enhanced teacher use of Maths-Whizz involves stronger integration of Maths-Whizz into maths teaching and learning (teacher direction of the Maths-Whizz curriculum, and/or using Maths-Whizz progress data to inform class teaching). These uses are expected to enhance pupil outcomes and increase teachers' use of adaptive teaching in maths. We anticipated that these longer-term uses and outcomes might be seen in some schools within the trial period, and this was explored through the IPE.

Figure 1: Maths-Whizz logic model



Issues

Recruitment to the trial was a bit slower than originally anticipated. In conversations with schools, it emerged that one of the blockers to recruitment was the exclusion of schools with mixed year groups from the trial. Our rationale for this criterion was that because of the interleaved design of the trial where successive year groups could be randomly allocated to different groups, allowing schools with mixed year groups could potentially cause contamination. However, given the slow pace of recruitment, it was decided (in agreement with Whizz Education and the EEF) that schools with mixed year groups could participate in the trial so long as their maths teaching was not mixed.

The number of pupils participating in the trial was ~10% higher than we expected. This was because of the recruitment of a number of multi-form entry schools, which put forward more than one class from each four of the year groups to the trial. Although this was not a major issue, it created operational challenges as we had to ensure that the online baseline assessment (Star Maths assessment and the Maths and Me survey) were set-up for this higher number of pupils within the short window between the end of recruitment and the start of baseline testing in the Summer Term.

Schools experienced some technological challenges with both the online Star Maths assessment and the Maths and Me survey. The NFER operations team had set-up unique Star Maths assessment accounts for each pupil for whom we received baseline data and created unique links to the Maths and Me survey for each pupil so that we could track completion of pupil surveys. Schools were provided with printable forms with the login details for both that were to be shared with the pupil ahead of the start of testing. Teachers reported issues with the internet connectivity that, in some instances, caused the Star Maths assessment to be paused part-way through. In such instances, teachers were contacted by the NFER operations team and provided with assistance in restarting the assessment. Issues with accessing the link to the Maths and Me survey were reported very early in the baseline testing window where pupils were, in some cases, locked out of the survey. These were quickly resolved by the operations team and updated troubleshooting guidance was sent to schools.

One of the issues we encountered with the baseline Star Maths assessment data was that there were some instances where there was more than one Star Maths assessment completed by an individual pupil. Of the 11,067 individual records in the baseline Star Maths assessment dataset where each record corresponds to an individual score, 541 records (4.5%) had this issue where more than one record mapped to the same pupil. When we consulted Renaissance Learning about this issue, they suggested that a possible explanation was that multiple pupils were completing the test using the same pupil login. This could have happened if the original pupil who logged in with the correct details did not fully log out of their account and subsequent pupils using the same device continued to use the same account rather than their unique login details. As this was the most likely explanation, in terms of data management, we used the first recorded score for the pupil and discarded subsequent scores. This meant that there was loss of some baseline data that we were unable to match to pupils. Pupils without baseline data were not included in the final analysis sample.

Evaluation objectives

The research questions (RQs) for the impact evaluation and IPE are provided below. Further detail on the evaluation design can be found in the study protocol (Sahasranaman *et al.*, 2025) and Statistical Analysis Plan (Schwendel, Welbourne and Sahasranaman, 2025).

Impact RQs

Primary RQ

RQ1: What is the overall impact of Maths-Whizz on the maths outcomes of children in Years 2 to 5?

Secondary RQs

RQ2: What is the impact of Maths-Whizz on each individual year group?

RQ3: What is the impact of Maths-Whizz on disadvantaged children as indicated by their FSM-eligibility status?

RQ4: Does the impact of Maths-Whizz vary for different genders?

RQ5: What is the impact of Maths-Whizz on children with low prior learning in maths?

RQ6: What is the impact of Maths-Whizz on children's motivation towards maths learning?

RQ7: What is the relationship between time spent (on tutor or practice modes) using Maths-Whizz (dosage) and improvement in maths outcomes?

RQ8: What is the relationship between Maths-Whizz estimated maths age and maths outcomes? (i.e. the relationship between the Maths-Whizz estimated maths age and the Star Maths scaled score at endline)

IPE RQs

IPE RQ1: To what extent was fidelity of implementation achieved?

IPE RQ2: What are the facilitators and barriers for implementing Maths-Whizz?

IPE RQ3: What are the perceived outcomes of Maths-Whizz?

IPE RQ4: What are mediators and moderators for Maths-Whizz outcomes?

IPE RQ5: What is 'business as usual', and to what extent was Maths-Whizz different?

IPE RQ6: What implementation characteristics improve maths outcomes for pupils?

IPE RQ7: What is the cost per pupil per year of Maths-Whizz?

Ethics and trial registration

An ethical review was undertaken as part of NFER's start-up meeting in January 2024 where consideration was given to consent and the impact of the research on trial participants (pupils and practitioners). The evaluation was conducted in accordance with the **NFER Code of Practice**. All of NFER's projects abide by its Code of Practice, which is in line with the Codes of Practice from the British Educational Research Association (BERA), the Market Research Association (MRA), and the Social Research Association (SRA), among others. NFER is committed to the highest ethical standards in all of its activities and ethical considerations are embedded in its detailed quality assurance processes.

Each participating school's headteacher provided their agreement to participate in the trial by signing the Memorandum of Understanding (MoU – see 'Further appendices' section) that outlined the responsibilities of all parties involved in the trial. NFER shared a parent letter and withdrawal form with schools to be sent to parents/carers of all pupils in all classes that schools intended to nominate for participation in the trial. Through the withdrawal form, parents/carers had the opportunity to withdraw their child from the evaluation and associated data processing at any stage of the trial.

A separate opt-in agreement process was used for the pupil focus groups and only applied to those selected to participate. Since pupils participating in the focus groups were only eight to ten-years-old, we could not assume that all pupils would have the capacity to provide fully informed consent to participate. In addition, as the focus groups involved audio recordings, it was especially important to ensure that parents/carers had the option to specifically agree to their child participating in this evaluation activity. We, therefore, provided parents/carers with a written information sheet containing full details about the focus group and what their child would be asked to do. Parents/carers were then asked to provide written opt-in agreement for their child to be invited to participate in the focus group, by returning a consent form to the school, who then passed this information on to the research team.

Pupil participation in the focus groups was voluntary, therefore, even if a parent/carer agreed for their child to participate, their child could still choose not to take part. Age-appropriate information about the focus groups was provided to pupils at the same time as parents/carers receive information about the focus groups to allow them to discuss participation together. The researchers also read this information to pupils at the beginning of the focus group to ensure pupils understood it and had the chance to ask any questions. If at any point a pupil decided that they would prefer not to participate, then they would be able to return to their class. Prior to beginning the focus group, the researchers agreed some ground rules for the group with the pupils and had a discussion with them about the types of scenarios in which we would need to break confidentiality, to ensure they fully understood what this means.

The trial was designed, conducted, and reported to Consolidated Standards of Reporting Trials (CONSORT) standards and registered on the International Standard Randomised Controlled Trial Number (ISRCTN) registry (**https://doi.org/10.1186/ISRCTN13444898**). The trial registry will also be updated with outcomes at the end of the project.

Data protection

Personal data was processed as part of this trial. All data gathered was held in accordance with the data protection framework created by the (UK Government, 2018) and the General Data Protection Regulation (GDPR) 2016/679 (GDPR, 2016) (European Union, 2016)and treated in the strictest confidence by the NFER, Whizz Education, and the EEF. NFER and Whizz Education are independent data controllers for the evaluation and the programme, respectively, for the duration of this trial. All relevant documents (including information sheets, MoU, withdrawal forms, consent form, privacy notices) are included in the 'Further appendices' section alongside this report.

The legal basis for processing personal data is covered by: GDPR Article 6 (1) (f) which states that:

> 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data' (GDPR, 2016)(European Union, 2016).

We have carried out a legitimate interest assessment, which demonstrates that the evaluation fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities). It also has broader societal benefits and will contribute to improving the lives of learners and teachers by providing evidence of the impact of virtual adaptive tutoring platforms on learning outcomes and classroom practice. Therefore, it is in our legitimate interest to process and analyse personal data for the administration of this RCT and the analysis of its impact on maths outcomes for pupils. Details of all data processed by NFER for this project are also recorded in the project's data log that is overseen by NFER's compliance officer.

NFER and Whizz Education signed a Data Sharing Agreement that governed the collection and sharing of personal data during this trial. This agreement included a description of the nature of the data being collected and how it would be shared, stored, protected, and reported by each party. In addition, Whizz Education also provided an MoU to schools, explaining the nature of the data being requested of schools, teachers, and pupils, how it would be collected, and how it would be passed to and shared with NFER. Two separate privacy notices, one for schools and another one for parents, are available **here**.

For the purposes of the trial, Whizz Education collected the names, roles, and contact details of a key contact person and the person signing the MoU when schools were recruited. They shared these data with NFER, who then contacted the key project contact person at participating schools to initiate data collection for the trial. NFER asked participating schools to share pupil data for all pupils in each participating class in Years 1, 2, 3, and 4 (in the Summer Term of 2024) and subsequently shared these data only for pupils in the intervention year groups with Whizz Education to enable them to create pupil accounts on the Maths-Whizz platform. NFER also collected data for class teachers in each participating class. This data was used to administer baseline and endpoint surveys to teachers and was also shared with Whizz Education to facilitate the coordination of training. Whizz Education will also share the names and contact details of the ESPs with NFER, so that NFER could contact them to attend and observe a sample of training sessions and conduct a focus group with ESPs. All personal data was shared via secure, password-protected data sharing portals.

NFER also collected pupil data from schools including names, date of birth, Unique Pupil Number (UPN), gender, FSM-eligibility status, current class name, and class name in next academic year for all pupils in participating classes in Years 1, 2, 3, and 4. The baseline and endpoint Star Maths assessment scores and the Maths and Me survey responses were collected for these pupils. In addition to this data, for pupils in the intervention year groups, NFER also received reports of

Maths-Whizz platform usage data. For all pupils in the trial, background data including gender and FSM-eligibility status was collected from the National Pupil Database (NPD). To obtain the information from the NPD, NFER provided the Data Sharing Team at the DfE with the names of the pupils, their dates of birth, and UPNs, allowing a match to NPD.

As part of the IPE, NFER conducted online surveys of teachers, senior leaders, and Maths-Whizz leads in each participating school, observations of training sessions, focus group with trainers (ESPs), and interviews with teachers. A small number of schools were also invited to participate in pupil focus groups. All NFER staff visiting schools had up-to-date Disclosure and Barring Service checks. All data gathered during interviews was stored securely. No names of individuals will be used in any report arising from this work.

Within three months of the end of project (i.e. publication of this report), NFER will send school and pupil data to the EEF's data archive partner. At this point, the EEF's data archive partner will keep a copy of the data and the EEF will become the data controller. NFER will retain personal data for one year after report publication in case there are any queries about the report. One year after the report publication (expected to be June 2026), all personal data will be securely deleted.

Project team

Table 4: Project team

Name	Organisation	Roles and responsibilities
Stephen Welbourne	NFER	Project director. Responsible for overall delivery and quality assurance of the trial
Aarti Sahasranaman	NFER	Project leader. Day-to-day management of the trial and overall delivery of trial design
Katherine Aston	NFER	IPE lead. Design and delivery of the IPE
Gustavo Lopes	NFER	IPE researcher. IPE data collection
Gemma Schwendel	NFER	Statistician
Kathryn Hurd	NFER	Research operations Lead. Test and schools administration lead, responsible for overseeing recruitment, school communications strategy, and testing
Jishi Jose	NFER	Operations manager. Day-to-day operations including preparation of recruitment documents, coordinating data collection, and point of contact for schools participating in the trial
Tom Dickinson	NFER	Operations researcher. Endpoint testing set-up and monitoring
Richard Marett	Whizz Education	Project director. Responsible for strategic oversight
Sarah Hawkes	Whizz Education	Project manager. Day-to-day operations including preparation of recruitment documents, coordinating recruitment processes/strategy, project management, and point of contact for the EEF/NFER
Emma Ringe	Whizz Education	Project implementation manager. Overall responsibility for recruitment of schools and implementation management during delivery period
Lindsay Brownsword	Whizz Education	School implementation manager. Monitoring school implementation at the project level, project implementation, and project management
Ray Douse	Whizz Education	Compliance director. Ensuring compliance with GDPR and other compliance measures

Methods

Trial design

Table 5: Trial design

Trial design, including number of arms		Two-arm, interleaved, four-level (pupil, class, year group, school) cluster randomised efficacy trial
Unit of randomisation		School
Stratification variable (s) (if applicable)		Not applicable
Primary outcome	Variable	Maths attainment
	Measure (instrument, scale, source)	Renaissance Learning Star Maths assessment Unified Scaled Score ranging from 600 – 1,400 (**www.renlearn.co.uk/**)
Secondary outcome(s)	Variable(s)	Motivation towards maths learning—mathematical self-perceptions and enjoyment of mathematics
	Measure(s) (instrument, scale, source)	Maths and Me survey consisting of two scales: Mathematical self-perceptions (scale = 8 – 40) Enjoyment of mathematics (scale = 10 – 50) (Adelson and McCoach, 2011)
Baseline for primary outcome	Variable	Maths attainment
	Measure (instrument, scale, source)	Renaissance Learning Star Maths assessment Unified Scaled Score ranging from 600 – 1,400 (**www.renlearn.co.uk/**)
Baseline for secondary outcome(s)	Variable	Motivation towards maths learning—mathematical self-perceptions and enjoyment of mathematics
	Measure (instrument, scale, source)	Maths and Me survey consisting of two scales: Mathematical self-perceptions (scale = 8 – 40) Enjoyment of mathematics (scale = 10 – 50) (Adelson and McCoach, 2011)

This was an efficacy trial to evaluate the impact of the Maths-Whizz virtual intelligent tutoring programme on the maths attainment of pupils from Years 2 to 5 over one academic year (2024/2025). In this two-arm school RCT, schools were randomised in a one to one ratio to two arms: schools in Arm 1 delivered Maths-Whizz to pupils in Years 3 and 5 and schools in Arm 2 delivered Maths-Whizz to pupils in Years 2 and 4. Pupils in year groups not receiving the programme in each arm served as controls for intervention pupils in the other arm. Therefore, all recruited schools delivered the Maths-Whizz programme, albeit in different year groups. The programme was delivered to pupils using a whole-class approach.

A total of 66 schools were recruited and provided pupil data. This included three pairs of infant (cover Reception to Year 2 and are typically feeder schools to their junior schools) and junior schools (cover Years 3 – 6), which were randomised as a single cluster to ensure that both the infant and junior schools in a pair were allocated to the same arm. Therefore, a total of 66 schools (or 63 clusters) were randomised.

All pupils in participating classes in all four of the year groups (Years 2, 3, 4, and 5) completed the baseline assessment prior to randomisation. The baseline assessment was completed in the Summer Term of 2024 when these pupils were in Years 1, 2, 3, or 4, respectively. The primary outcome for this trial was pupils' maths attainment as measured by the Star Maths assessment scaled scores. The secondary outcome was pupils' motivation towards maths learning measured by the Maths and Me survey. This validated survey consists of two scales: i) mathematical self-perceptions; and ii) enjoyment of mathematics.

Participant selection

Maths-Whizz was delivered to the whole class rather than to individuals or subsets of pupils. Pupils in Years 2, 3, 4, and 5 were eligible to receive the programme. The teachers for each class were responsible for managing the delivery of the programme and participated in the IPE, alongside school leaders and maths ambassadors.

Whizz Education was responsible for the recruitment of schools to this trial. They achieved this target through a combination of contacting schools on their mailing list and reaching out to new schools not on their contact list. The EEF and NFER supported their recruitment efforts by promoting the trial through their newsletters and social media channels. Whizz Education offered webinars during the recruitment period to provide information on the trial to schools interested in participating. Interested schools completed an online Expression of Interest (EOI) form, which helped Whizz Education ascertain schools' eligibility to participate in the trial. School eligibility criteria are set out below. Whizz Education followed up with eligible schools via a phone call to confirm eligibility and clarify details such as the number of classes in each year group that would participate in the trial. Eligible schools were then sent the school information sheet and MoU. Schools signed up to the trial by the headteacher signing the MoU and providing the name of a key project contact who would act as the coordinator of the trial in the school. Once a school signed up, Whizz Education shared the school name and details of the headteacher and key project contact with NFER. NFER then got in touch with the key project contact to initiate data collection and other trial-related activities.

School eligibility

Schools were eligible for the trial if they met the following criteria:

- Were a state-funded infant/junior/primary school in England with children in Years 2, 3, 4, and 5 as of 01 September 2024.
- Did not have mixed year maths teaching even if the school had mixed year groups (i.e. children who are from more than one year group within the same class) in any of Years 2, 3, 4, and 5 in the 2024/2025 academic year.
- Had not implemented Maths-Whizz or any Whizz associated services in the 2023/2024 academic year.
- Were not participating in another trial funded by the EEF in the 2024/2025 academic year.
- Had a suitable level of IT provision including devices such as tablets, laptops, desktops and sufficient Wi-Fi™ capacity (0.5 Mbps internet connection per concurrent pupil) (see Maths-Whizz minimum requirements).
- Agreed to contribute **£500** towards the cost of the programme (usual cost is £1,875 per school plus £25 per pupil per year).

Pupil eligibility

There were no pupil eligibility criteria as the Maths-Whizz programme was accessed by the whole class. Pupils with severe visual impairments were not able to access the programme.

Outcome measures

Baseline measures

The baseline measure for all impact RQs except RQ6 was the commercially available Star Maths Unified Scaled Score, which puts all pupils, regardless of their year group, onto the same scale. This online adaptive test, which is expected to take pupils 20–25 minutes to complete, was administered in schools by teachers in the Summer Term prior to commencement of the Maths-Whizz intervention. The pupils taking part in this trial took, on average, 12.3 minutes to complete both their baseline and endline assessments.

The baseline measure for RQ6 was the Maths and Me survey (Adelson and McCoach, 2011), adapted to make it suitable for administration in UK schools (this simply involved replacing the word ‘math’ with ‘maths’ throughout). This online survey, which takes around five minutes to complete, was also administered in schools by teachers.

To minimise the burden on pupils, teachers were instructed to administer the Maths and Me survey following a short break of 10–15 minutes after the completion of the Star Maths assessment. The entire baseline testing period for each pupil (including the Star Maths assessment and Maths and Me survey) was expected to last around 45 minutes.

We provided detailed baseline test administration and troubleshooting guidance to schools, which also suggested that schools complete the assessments with Year 4 first, as these were the oldest children participating and therefore, likely to be more independent and require less support than the younger pupils. Carrying out the baseline activities with Year 4 would enable teachers and schools to become familiar with the format, before proceeding with the younger year groups.

Primary outcome

Similar to baseline, the primary outcome measure for all impact RQs except RQ6 was the Star Maths Unified Scaled Score. This endpoint online adaptive Star Maths assessment was overseen by independent NFER test administrators and was administered to all pupils in participating classes in summer 2025. NFER test administrators underwent training that covered the administration and troubleshooting of the Star Maths online assessment.

Secondary outcomes

The secondary outcome (RQ6) was explored using the Maths and Me survey (see Appendix E), also overseen by independent NFER test administrators. The suggested format of endpoint testing was similar to that of baseline testing, i.e., where possible, older year groups were tested first and pupils were allowed a short break of 10–15 minutes between the Star Maths assessment and Maths and Me survey. The survey was administered at both baseline and endpoint. However, due to concerns about the youngest children not having sufficient reading and comprehension skills to be able to access the survey fully, it was only administered to those who were in Years 3, 4, and 5 at endpoint (i.e. Years 2, 3, and 4 at baseline).

Each question was answered via a 5-point Likert scale (‘Not at all like me’, ‘Not like me’, ‘Sort of like me’, ‘Like me’, ‘Very like me’) and the subscales calculated by summing the appropriate responses. Where questions were ‘negative’, they were recoded so that ‘Not at all like me’ scores 5 and ‘Very like me’ scores 1. Responses from the survey were used to construct two scales: mathematical self-perceptions; and enjoyment of maths.

Sample size

For the sample size calculations at design stage, analysts based these estimates on parameters from the EEF latest guidance on power calculations (Singh *et al.,* 2023), taking the median values based on NPD data between 2012 and 2019. We were not able to find any data looking at how the intracluster correlation coefficient (ICC) for primary maths would divide if separated into year group and school levels. We assumed that the vast majority of variance from these two levels would be accounted for at the year group level as these children will be taught the same curriculum usually by the same teacher. Considering the absence of strong evidence supporting this assumption, ICC was split relatively conservatively, with one-third at the school level and the remainder at year group level. We also assumed a pupil-level attrition of 15% in line with our experience in similar trials. The assumptions about average class size, the average number of classes per year group and the proportion of pupils who have ever been eligible for FSM were based on the latest data available at the time from DfE (DfE, 2023). The sample size was specified at design stage so that the evaluation would be powered for a minimum detectable effect size (MDES) of 0.2 for the analyses of individual year groups, which was deemed to be the limiting factor. By powering the trial for the individual year groups analysis, the trial was well powered for all pupils by default.

The sample size calculations at analysis stage used parameters derived from the pupils and schools who took part in the trial. All sample size calculations and their parameters are set out in Table 13.

Randomisation

Randomisation occurred at the school level, with schools allocated to Arm 1 or Arm 2. Schools in Arm 1 delivered the intervention to classes in Years 3 and 5, while schools in Arm 2 delivered it to Years 2 and 4. In each arm, classes in the other year groups served as controls. No stratification variables were used during randomisation. Three infant–junior school pairs participated in the trial. These pairs were treated as single clusters for both randomisation and subsequent analyses to ensure that they were randomised to the same group allocation. In total, 66 schools (63 clusters) were randomised. Randomisation took place after baseline assessments were completed. R version 4.5.0 was used to implement the randomisation procedure (see 'Randomisation code' in the Statistical Analysis Plan; Schwendel, Welbourne, and Sahasranaman, 2025).

The analyst was not blinded during randomisation. To minimise potential bias, the following steps were taken:

- **Pre-specified randomisation algorithm.** The randomisation procedure was fully scripted in R and executed without manual intervention, ensuring allocation was determined by a reproducible algorithm.

- **Independent verification.** The randomisation code and outputs were checked by a second analyst to confirm correctness and adherence to the protocol.

The number of schools, classes, and pupils in each arm and then in the control and intervention groups are shown in Table 6 and Table 7, respectively.

Table 6: Randomisation outcome[a] by arm

	Arm 1	Arm 2	Total
Number of schools (%)	33 (50%)	33 (50%)	66
Number of clusters (%)	31 (49%)	32 (51%)	63
Number of classes (%)	208 (47%)	236 (53%)	444
Number of pupils (%)	5,314 (51%)	5,055 (49%)	10,369

[a]*Note, when the randomisation outcome information was sent to schools, the outcomes were inadvertently reversed due to a mail merge error (i.e. those originally randomised to Arm 1 were told they had been assigned to Arm 2 and vice versa). Following consultation with the EEF and the developer it was decided to adopt the reversed assignments to avoid confusing the schools and placing extra burden on them. This reversal will have no impact on the results. The figures shown in the table reflect these reversed assignments.*

Table 7: Randomisation outcome[a] by intervention and control groups

	Control	Intervention	Total
Number of schools (%)	66	66	66
Number of classes (%)	223 (50%)	221 (50%)	444
Number of pupils (%)	5,136 (49%)	5,233 (51%)	10,369

[a]*Note, when the randomisation outcome information was sent to schools, the outcomes were inadvertently reversed due to a mail merge error (i.e. those originally randomised to Arm 1 were told they had been assigned to Arm 2 and vice versa). Following consultation with the EEF and the developer it was decided to adopt the reversed assignments to avoid confusing the schools and placing extra burden on them. This reversal will have no impact on the results. The figures shown in the table reflect these reversed assignments.*

Statistical analysis

The primary and secondary analyses followed the EEF 2022 guidelines(EEF, 2022b) and assumed intention-to-treat (ITT).

Primary analysis

A multilevel random intercepts model with four levels (school, year group, class, and pupil) was used to account for cluster randomisation. The primary analysis investigated whether a class's participation in the Maths-Whizz programme had an effect on its pupils' maths attainment at the end of the 2024/2025 academic year. This was determined by fitting a model with maths attainment at follow-up, as measured by Star Maths Unified Scaled Score, as the dependent variable. The model was originally defined within the Statistical Analysis Plan (Schwendel, Welbourne, and Sahasranaman, 2025) as:

$$StarMaths_{ijkl} = \beta_0 + u_{0l} + v_{0kl} + w_{0jkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \varepsilon_{ijkl}$$

Where $StarMaths_{ijkl}$ is the endpoint Star Maths Unified Scaled Score for pupil i in class j in year group k in school l, u_{0l} is the intercept for school l, v_{0kl} is the intercept for year group k in school l, w_{0jkl} is the intercept for class j in year group k in school l, $intervention_{kl}$ is the year group level intervention/control dummy variable for year group k in school l, and $BLStarMaths_{ijkl}$ is the baseline Star Maths Unified Scaled Score for pupil i in class j in year group k in school l. ε_{ijkl} is the residual error term for pupil i in class j in year group k in school l.

However, when carrying out the analysis, attempting to run the model as specified in the Statistical Analysis Plan generated a 'boundary: singular fit' error(Schwendel, Welbourne and Sahasranaman, 2025). This was a result of pupils of different year groups being measured on the same scale, with younger pupils tending to score lower on average, leaving very little residual variation within year groups for the random effects to capture. Year group therefore, needed to be considered as a fixed effect as well as a random effect to account for this. Note that the random effect of year group is fundamentally different from the fixed effect as the random effect will be unique to each year group within each school whereas the fixed effect is common across all schools. Therefore, the revised model was:

$$StarMaths_{ijkl} = \beta_0 + u_{0l} + v_{0kl} + w_{0jkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \beta_3 YearGroup_{ijkl} + \varepsilon_{ijkl}$$

Where $StarMaths_{ijkl}$ is the endpoint Star Maths Unified Scaled Score for pupil i in class j in year group k in school l, u_{0l} is the intercept for school l, v_{0kl} is the intercept for year group k in school l, w_{0jkl} is the intercept for class j in year group k in school l, $intervention_{kl}$ is the year group level intervention/control dummy variable for year group k in school l, $BLStarMaths_{ijkl}$ is the baseline Star Maths Unified Scaled Score for pupil i in class j in year group k in school l and $YearGroup_{ijkl}$ is the ordinal year group variable (with four levels) for pupil i in class j in year group k in school l.. ε_{ijkl} is the residual error term for pupil i in class j in year group k in school l.

The residuals at the four levels were assumed to be independently normally distributed:

$$u_{0l} \sim N(0, \sigma_{u0}^2)$$

$$v_{0kl} \sim N(0, \sigma_{v0}^2)$$

$$w_{0jkl} \sim N(0, \sigma_{w0}^2)$$

$$e_{ijkl} \sim N(0, \sigma_e^2)$$

The model was run in R (version 4.5.0) using the package 'lme4' (Bates *et al.*, 2015).

Secondary analysis

Subgroup analyses (RQ2–RQ5)

A number of subgroup analyses were conducted that used the primary outcome measure and corresponding baseline measure. These subgroup analyses sought to address RQ2–RQ5, which focused on the impact of the intervention on the primary outcome analysis. Each is outlined in more detail below. All models were run in R version 4.5.0 using the 'lme4' package.

<u>Individual year group (RQ2)</u>

Two analyses were carried out for this subgroup analysis. The first ran a model with an interaction term between year group and group allocation, and the second ran a primary outcome model on each year group individually. Running separate models for each year group in this way removes year group from each analysis, thus reducing the levels to three: pupil, class, and school. However, year group was retained as a random effect in the interaction model. This approach differs from the model specified in the Statistical Analysis Plan (Schwendel, Welbourne and Sahasranaman, 2025), where year group was not included in the models for individual year groups. The deviation reflects a corresponding change made to the primary outcome analysis (see above).

The four-level random intercepts model with interaction term is given by:

$$StarMaths_{ijkl} = \beta_0 + u_{0l} + v_{0kl} + w_{0jkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \beta_3 YearGroup_{ijkl} + \beta_4 YearGroup_{ijkl} * intervention_{kl} + \varepsilon_{ijkl}$$

The three-level random intercepts model for the individual year group analyses is given by:

$$StarMaths_{ijl} = \beta_0 + u_{0l} + v_{0jl} + \beta_1 intervention_{jl} + \beta_2 BLStarMaths_{ijl} + \varepsilon_{ijl}$$

In each, variable definitions and assumptions about the distribution of the residuals are as per the primary outcome model.

FSM-eligibility status (RQ3)

In a similar manner and conforming to the EEF statistical analysis guidance (EEF, 2022), two models were run to investigate possible differential effects of Maths-Whizz on the maths attainment of children eligible for FSM. The first ran a model with an interaction term (between FSM and group allocation), and the second ran a primary outcome model on just the FSM-eligible pupils. EVERFSM_6_P collected from the NPD was used as the FSM identifier for this analysis.

The four-level random intercepts model with interaction term is given by:

$$StarMaths_{ijkl} = \beta_0 + u_{ol} + v_{okl} + w_{ojkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \beta_3 YearGroup_{ijkl} + \beta_4 FSM_{ijkl} + \beta_5 FSM_{ijkl} * intervention_{kl} + \varepsilon_{ijkl}$$

Where FSM_{ijkl} is a dummy variable indicating the FSM eligibility of pupil i in class j in year group k in school l. All other variables and the residuals at the four levels are as those defined for the primary outcome analysis.

Gender subgroup (RQ4)

Similar to the FSM subgroup analysis, a model was run with an interaction term between gender and group allocation, alongside separate primary outcome models for boys and girls. The four-level random intercepts model with interaction term is given by:

$$StarMaths_{ijkl} = \beta_0 + u_{ol} + v_{okl} + w_{ojkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \beta_3 YearGroup_{ijkl} + \beta_4 Gender_{ijkl} + \beta_5 Gender_{ijkl} * intervention_{kl} + \varepsilon_{ijkl}$$

$Gender_{ijkl}$ is the gender of pupil i in class j in year group k in school l. All other variables and the residuals at the four levels are as those defined for the primary outcome analysis.

Low prior attainment subgroup (RQ5)

This analysis explored the impact of Maths-Whizz on maths attainment and whether it varied by prior maths learning. The baseline Star Maths age-standardised score was used to define a flag for low prior maths attainment, where 1 = low prior attainment and 0 = not low prior attainment. A pupil was defined as having low prior attainment if their age-standardised score was less than 85, as Renaissance Learning classify a pupil with an age-standardised score of less than 85 as requiring intervention. Around 16% of pupils who completed the baseline were flagged as having low prior attainment using this definition. A subgroup analysis was then conducted along the same lines as that for other subgroups. Two models were run: one that included an interaction term between low prior attainment and group allocation; and a second that used a primary outcome model only on pupils with low prior attainment.

The four-level random intercepts model with interaction term is given by:

$$StarMaths_{ijkl} = \beta_0 + u_{ol} + v_{okl} + w_{ojkl} + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \beta_3 YearGroup_{ijkl} + \beta_4 Attain_{ijkl} + \beta_5 Attain_{ijkl} * intervention_{kl} + \varepsilon_{ijkl}$$

$Attain_{ijkl}$ is the dummy variable representing prior attainment (1 = low prior attainment; 0 = not low prior attainment) of pupil i in class j in year group k in school l. All other variables, and the residuals at the four levels are as those defined for the primary outcome analysis.

Secondary outcome analysis (RQ6)

For this analysis, which was undertaken to address RQ6, the scores for the two subscales of the Maths and Me survey ('mathematical self-perceptions' and 'enjoyment of mathematics') were calculated as outlined in the 'Outcome measures' section earlier in the report. They were then used as the dependent variables in two separate multilevel models. As per the primary outcome analysis, each model assumed the data was clustered at four levels, with class, year group, and school treated as random effects. However, in contrast to the primary outcome analysis, year group was not treated a fixed effect, as Maths and Me scores were on a common scale for all pupils irrespective of year group.

The two four-level random intercepts models are given by:

$$SelfPercept_{ijkl} = \beta_0 + u_{0l} + v_{0kl} + w_{0jkl} + \beta_1 intervention_{kl} + \beta_2 BLSelfPercept_{ijkl} + \varepsilon_{ijkl}$$

$$Enjoy_{ijkl} = \beta_0 + u_{0l} + v_{0kl} + w_{0jkl} + \beta_1 intervention_{kl} + \beta_2 BLEnjoy_{ijkl} + \varepsilon_{ijkl}$$

Where $SelfPercept_{ijkl}$ is the endpoint Self-Perception subscale score for pupil i in class j in year group k in school l, u_{0l} is the intercept for school l, v_{0kl} is the intercept for year group k in school l, w_{0jkl} is the intercept for class j in year group k in school l, $intervention_{kl}$ is the year group level intervention/control dummy variable for year group k in school l, $BLSelfPercept_{ijkl}$ is the baseline Self-Perception subscale score for pupil i in class j in year group k in school l, $Enjoy_{ijkl}$ is the endpoint Enjoyment subscale score for pupil i in class j in year group k in school l, and $BLEnjoy_{ijkl}$ is the baseline Enjoyment subscale score for pupil i in class j in year group k in school l. ε_{ijkl} is the residual error term for pupil i in class j in year group k in school l.

The residuals at the four levels are as those defined for the primary outcome analysis, and the analysis was carried out within R (version 4.5.0), with modelling performed using the lme4 package.

Analysis in the presence of non-compliance

Compliance for this trial has been measured using data from the Maths-Whizz platform and is assessed at the class level in order to help understand whether Maths-Whizz would, on average, benefit *all* pupils if a school were to implement it as expected. It is highly probable that, during the trial, there were individual pupils who did not access the Maths-Whizz platform as expected. However, by taking a class-level approach to measuring compliance, these pupils were not excluded from the analysis, and the impact of compliance for all pupils within the school has been assessed. We considered the proportion of classes and pupils reaching the 45-minute threshold referred to in the logic model, as well as a 30-minute threshold as a lower expectation. The lower threshold of 30 minutes was chosen because the compliance analysis focused only on usage in the tutor mode. Weekly tutor mode usage was assessed during term-time only, since compliance was measured at the class level and therefore, did not incorporate home usage.

- For this trial, a class has been deemed compliant if: Pupils in the class spent at least 30 minutes tutor usage per week, on average, in the total period considered (term-time only/excluding the holidays).

OR

- Pupils in the class attempted three tutor lessons per week (pass, static, or fail), on average, in the total period considered (term-time only/excluding the holidays).

The linear effect of this compliance indicator on the primary outcome measure has been explored using a complier average causal effect (CACE) estimate that has been obtained using instrumental variable (IV) modelling as prescribed by the EEF statistical analysis guidelines (EEF, 2022). The IV regression has been run via a two-stage least squares model with group allocation as the instrumental variable for compliance and is given by:

Stage one:

$$\widehat{Compliance}_{jkl} = \beta_0 + \beta_1 intervention_{kl} + \beta_2 BLStarMaths_{ijkl} + \varepsilon_{ijkl}$$

Stage two:

$$StarMaths_{ijkl} = \beta_0 + \beta_1 \widehat{Compliance}_{jkl} + \beta_2 BLStarMaths_{ijkl} + \varepsilon_{ijkl}$$

Where $\widehat{Compliance}_{jkl}$ is the estimated compliance measure for class j in year group k in school l obtained from the first stage of the two-stage regression, $intervention_{kl}$ is the year group level intervention/control dummy variable for year group k in school l, $BLStarMaths_{ijkl}$ is the baseline Star Maths Unified Scaled Score for pupil i in class j in year group k in school l and $StarMaths_{ijkl}$ is the endpoint Star Maths Unified Scaled Score for pupil i in class j in year group k in school l. The residuals in both stages are assumed to be independently normally distributed:

$$e_{ijkl} \sim N(0, \sigma^2)$$

The model was fitted using the function ivreg from the R package 'ivreg' (Fox *et al.*, 2026) and the estimation of causal effects has been completed using the functions contained in the 'ivpack' (Jiang and Small, 2014) package. The analyses were run in R (version 4.5.0).

Missing data analysis

Due to the study design (only pupils that had completed baseline Star Maths assessment completed it at endpoint), there were no missing baseline Star Maths scores. However, 323 pupils had missing class information on their pupil Masterfile record. This was addressed by first, collecting class name information from platform data where available, and second, using year group as a proxy for class name. For year groups with fewer than 30 pupils, this likely represented a single class and was therefore, not problematic. For 221 pupils across three schools with more than 30 pupils per year group, this approach likely reduced the number of clusters, but we consider this unlikely to bias results. Following this, two pupils remained without class information and were excluded from all analyses. Around 13% of pupils had missing primary outcome data, including one setting that withdrew from the trial and did not complete endpoint testing. As missingness exceeded 5%, we followed the EEF statistical guidance (EEF, 2022) to address it.

First, we examined patterns of missingness using a logistic regression model where the outcome was 'missingness'. Covariates included baseline Star Maths Unified Scaled Score, year group, and randomisation group indicators, as well as additional variables hypothesised to relate to missingness but not included in the primary analysis. At the school level, these were FSM eligibility (schools were grouped into quintiles according to the proportion of FSM-eligible pupils), urban/rural status, latest Office for Standards in Education, Children's Services and Skills (Ofsted) rating, and region. At the pupil level, these were FSM eligibility, gender, year group, attendance rate, and low prior attainment flag.

Results indicated that low prior attainment, FSM status, and absence rate were significantly associated with missingness (see Table 26). Two of these—prior attainment and FSM eligibility—were fully observed, while absence rate was missing for fewer than ten pupils. Following the EEF guidance (EEF, 2022), we conducted a sensitivity analysis by re-running the primary outcome model with pupil-level FSM status, low prior attainment, and absence rate included. Including these variables in the complete-case analysis (see Table 26) showed no impact on the primary outcome. Because these predictors did not act as substantive confounders and the effect estimate remained stable, multiple imputation was unlikely to provide additional benefit. We therefore, judged complete-case analysis to be adequate and proportionate in this context.

Additional analyses and robustness checks

Dosage analysis (RQ7)

The impact of dosage of the programme on pupils' maths attainment was explored by analysing usage data (time spent by pupils in tutor mode or time spent by pupils in practice mode) from the Maths-Whizz platform. In tutor mode, pupils learn new content, whereas in practice mode, pupils review, practice, and consolidate content they have previously covered.

For the dosage analysis, time spent in tutor mode and practice mode was analysed separately at the pupil level as per the Statistical Analysis Plan (Schwendel, Welbourne and Sahasranaman, 2025). The platform calculated time in each mode as the mean number of minutes per week over the full trial period, including school holidays. These figures were used without adjustment for holiday periods in the dosage analysis as some pupils were given home access and could potentially have accessed Maths-Whizz during holidays.

For the IPE analysis, overall Quality Learning Time was calculated within the platform by summing the weekly mean time spent in tutor mode and practice mode. These figures were presented: i) across the whole trial period; and ii) with the time adjusted to exclude an assumed eight weeks of holiday. Class level means were then derived by taking the mean of these adjusted pupil-level means.

A linear model was used that set the endpoint Star Maths Unified Scaled Score as the dependent variable with two dosage covariates. The first measured the amount of time spent in tutor mode, while the second measured the amount of time spent in practice mode. Additionally, the degree of correlation between the two was assessed. If the two had been sufficiently correlated, they would have been combined to give a single dosage predictor. However, this proved to not be the case, so the two covariates were treated separately.

The four-level random intercepts model for the analysis is given by:

$$StarMaths_{ijkl} = \beta_0 + u_{ol} + v_{okl} + w_{ojkl} + \beta_1 tutor_{ijkl} + \beta_2 practice_{ijkl} + \beta_3 BLStarMaths_{ijkl} + \varepsilon_{ijkl}$$

Where $StarMaths_{ijkl}$ is the endpoint Star Maths Unified Scaled Score for pupil i in class j in year group k in school l, u_{0l} is the intercept for school l, v_{0kl} is the intercept for year group k in school l, w_{0jkl} is the intercept for class j in year group k in school l ,$tutor_{ijkl}$ is the time spent in tutor mode for pupil i in class j in year group k in school l, $practice_{ijkl}$ is the time spent in practice mode for pupil i in class j in year group k in school l, and $BLStarMaths_{ijkl}$ is the baseline Star Maths Unified Scaled Score for pupil i in class j in year group k in school l. The residuals at the four levels are as those defined for the primary outcome analysis.

Correlation analysis (RQ8)

The Maths-Whizz tutor prepares a differentiated learning journey for each pupil based on their maths age, which is calculated within the Maths-Whizz platform. This analysis explored the relationship between Maths-Whizz calculated maths age and maths outcomes as measured by the Star Maths assessment and calculated by the correlation coefficient between pupils' endpoint maths age and endpoint Star Maths Unified Scaled Scores.

Estimation of effect sizes

As specified in the EEF statistical analysis guidelines (EEF, 2022), the primary outcome results estimated via multilevel regression models are reported as Hedges' g. In a single year group trial, g would ordinarily be calculated as:

$$g = \frac{\bar{o}_i - \bar{o}_c}{s^*}$$

where $\bar{o}_i - \bar{o}_c$ is the adjusted mean difference between the intervention and control groups (i.e. the treatment coefficient from the full regression model), and s^* is the pooled *unconditional* standard deviation (SD) across the two groups.

However, this standard two-group pooling approach is not appropriate for the structure of the present trial, which includes four distinct year groups. Using a pooled SD that ignores year group structure would combine outcomes from children who differ systematically by age, because variability across four of the year groups will naturally be larger than variability within a single year group. This inflates the denominator of the effect size, inevitably producing an effect size that is smaller than the intervention effect within each year group, misrepresenting the magnitude of impact for a multi-year design.

To avoid this and ensure comparability across year groups, the effect size has therefore, been calculated in a way that fully controls for between-year heterogeneity:

- The numerator, $\bar{o}_i - \bar{o}_c$, is the adjusted difference in means calculated using Unified Scaled Scores, which place all pupils on a common developmental scale.
- The denominator is an extended pooled SD computed across eight groups (intervention and control within each of Years 2, 3, 4, and 5):

$$s^* = \sqrt{\frac{(N_{i2}-1)s_{i2}^2 + (N_{c2}-1)s_{c2}^2 + (N_{i3}-1)s_{i3}^2 + (N_{c3}-1)s_{c3}^2 + (N_{i4}-1)s_{i4}^2 + (N_{c4}-1)s_{c4}^2 + (N_{i5}-1)s_{i5}^2 + (N_{c5}-1)s_{c5}^2}{N_{i2} + N_{c2} + N_{i3} + N_{c3} + N_{i4} + N_{c4} + N_{i5} + N_{c5} - 8}}$$

where N_{iy} and N_{cy} denote the intervention and control sample sizes in year group y, and the corresponding s values represent the within-group SDs.

This approach ensures that each year group contributes appropriately to the pooled SD, preserving the integrity of within-year variation while preventing systematic age-related *differences in score distributions across year groups* from distorting the effect size. As shown in Table 18, the means and standard errors of the raw Unified Scaled Scores differ across year groups in ways that reflect expected developmental progression. Therefore, pooling variability across only two broad groups (intervention vs control) conflates this between-year variation with within-group variation. Using year-specific SDs avoids this distortion and provides a more comparable basis for calculating effect sizes.

An alternative approach would be to calculate the effect size using age-standardised scores, with the pooled SD calculated over only two groups (intervention and control). While this would adjust for between-year differences, it would introduce a substantive inconsistency: the effect size would be derived from age-standardised scores, whereas the descriptive group means reported in the tables would be based on Unified Scaled Scores. This mismatch complicates interpretation and reduces transparency. For this reason, all reported effect sizes based on the primary outcome measure use Unified Scaled Scores with the eight-group pooled SD described above. The only exception is the year group subgroup analyses that are conducted within a single year group only. For these, the standard two-group pooled SD is used.

To demonstrate that the chosen approach is appropriate and does not materially distort the conclusions, two supplementary effect sizes have also been calculated and presented for the primary outcome analysis:

1. Effect size using age-standardised scores, with the numerator computed as the intervention–control difference in age-standardised means, and the denominator computed as the standard two-group pooled SD:

$$s^* = \sqrt{\frac{(N_i - 1)s_i^2 + (N_c - 1)s_c^2}{N_i + N_c - 2}}$$

where s_i and s_c are the group-specific SDs for the age-standardised outcome.

2. Effect size using age-standardised scores with the eight-group pooled SD, paralleling the primary method but replacing scaled scores with age-standardised ones.

Together, these sensitivity analyses provide reassurance that the primary method appropriately addresses between-year variability, produces interpretable effect sizes, and avoids the bias that would result from a conventional two-group pooled SD.

Confidence intervals (CIs) for each effect size have been computed by multiplying the standard errors of the intervention group by the left-tailed inverse of the student's t-distribution with a probability of 2.5% and the number of degrees of freedom associated to the intervention group. The CIs for the standard errors have been converted to effect size CIs using the same formula as the effect sizes themselves.

Estimation of ICC

For the primary outcome analysis, ICCs have been calculated at three levels:

$$ICC_{class} = \frac{\sigma_j^2}{\sigma_j^2 + \sigma_k^2 + \sigma_l^2 + \alpha}$$

$$ICC_{yeargroup} = \frac{\sigma_k^2}{\sigma_j^2 + \sigma_k^2 + \sigma_l^2 + \alpha}$$

$$ICC_{school} = \frac{\sigma_l^2}{\sigma_j^2 + \sigma_k^2 + \sigma_l^2 + \alpha}$$

Where σ_j^2 is the between-class variance, σ_k^2 is the between-year group variance, σ_l^2 is the between-school variance, and α is the between-pupil variance.

Pre-test ICCs have been derived from random intercepts models with the appropriate number of levels and no covariates, and post-test ICCs were derived from the primary outcome model (using Star Maths Unified Scaled Scores) described above.

IPE

The IPE followed the **EEF's updated IPE guidance** and complements the impact evaluation by exploring *how* and *in what circumstances* Maths-Whizz impacts teachers and pupils.

As this was an efficacy trial, a key focus was implementation fidelity. Given the complexity of the theory of change (ToC), we aimed to identify how key implementation characteristics contributed to key outcomes, in order to refine the ToC and inform potential scale-up. We compared FSM-eligible pupils with non-eligible pupils to describe any differences in implementation, experience, and perceived outcomes for FSM pupils.

Considering how Maths-Whizz is implemented, our IPE tested the logic model's hypothesis about the leader/teacher support (activities, inputs, and outputs) needed to facilitate weekly implementation of Maths-Whizz tutoring (IPE RQ1). Through the IPE evaluation, we have identified key support needed for implementation, and barriers or challenges encountered (IPE RQ2). We calculated the cost to schools of delivering Maths-Whizz (IPE RQ7). Considering perceived impact, we checked the pupil outputs and outcomes in the logic model by describing teachers' and pupils' perceptions of impact and the key features and mechanisms which facilitated this impact (IPE RQ3). Considering variation in approaches to using Maths-Whizz, we described the extent of integrated use of Maths-Whizz in maths teaching (shown as blue boxes in the logic model) (IPE RQ4). We used these findings to describe the perceived importance and prevalence of different pathways to impact (IPE RQ5).

To contextualise our impact evaluation's comparison of the intervention and control groups, we described 'business as usual' for both arms, monitoring this during the trial for the control group. This allowed us to assess programme differentiation for the intervention group.

IPE RQs are described below.

IPE RQ1: To what extent was fidelity of implementation achieved? [Fidelity, Quality, Adaptation]

- Did school leaders/teachers attend the relevant onboarding and reviews?
- Was Maths-Whizz used for at least 30 minutes per week for each class?
- Was Maths-Whizz used for at least 30 minutes per week for each pupil?
- How did implementation vary across and within schools, including the additionality of Math-Whizz time and planned home usage?
- Was there any difference in implementation in schools with different proportions of FSM pupils?
- Were there any differences in Maths-Whizz use between FSM and non-FSM pupils?
- Where fidelity was not achieved, why, and what adaptations were made?

A key fidelity measure was the weekly minutes of usage recorded in Maths-Whizz platform data (including the 'tutor' mode for new learning, and the 'practice' mode for revision). We explored the distribution of average usage time at class level and pupil level. We considered the proportion of classes and pupils reaching the 45-minute threshold referred to in the logic model, as well as a 30-minute threshold as a lower expectation. Our impact evaluation assessed compliance based on time spent in 'tutor' mode only and assessed the impact of dosage on maths outcomes.

IPE RQ2: What are the facilitators and barriers for implementing Maths-Whizz? [Context, Responsiveness]

Based on the logic model and the Intervention Delivery and Evaluation Analysis (IDEA) workshop, we identified three key implementation factors which informed our RQs: quality of IT infrastructure; teacher/leader engagement and efficacy (supported by Maths-Whizz implementation inputs); and how Maths-Whizz affects workload (EEF, 2019; DfE, 2022a).

- What facilitated and hindered schools in using Maths-Whizz? (e.g. quality of IT infrastructure, devices, spaces, timetabling, teacher confidence, home use).
- Were leaders, teachers, pupils, and parents engaged? What facilitated and hindered this (e.g. Maths-Whizz support)?
- Were any barriers/facilitators experienced specifically or differentially by FSM pupils?
- How has Maths-Whizz affected teacher workload?
- Overall, what worked well and less well, why?

IPE RQ3: What are the perceived outcomes of Maths-Whizz? [Perceived impact]

- What are perceived outcomes for pupils (e.g. attainment, motivation, enjoyment, confidence, self-efficacy, maths anxiety)?
- What are perceived outcomes for teachers and schools (e.g. digital literacy, formative assessment, adaptive teaching, workload)?
- Are there any unintended consequences for pupils, teachers, or schools?

IPE RQ4: What are mediators and moderators for Maths-Whizz outcomes? [Mediators, Moderators]

Based on the logic model and the IDEA workshop, we identified three pathways which may contribute to improved maths outcomes. First, learning within Maths-Whizz may directly improve maths outcomes. Second, using Maths-Whizz may improve pupils' engagement with maths teaching, mediating improved maths outcomes (e.g. filling learning gaps improves pupils' access to class teaching). Third, changes in teacher practice (e.g. using Maths-Whizz data to tailor class teaching) may mediate improved maths outcomes.

- To what extent are perceived pupil engagement with maths teaching and changes in teacher practice associated with enhanced Maths-Whizz outcomes?
- Do teachers perceive any differences in Maths-Whizz usage or outcomes for different pupils?
- Were there any differences in perceived outcomes for FSM pupils?

Our impact evaluation assessed the impact of Maths-Whizz for pupils with different characteristics (e.g. gender, levels of prior attainment, FSM eligibility).

IPE RQ5: What is 'business as usual', and to what extent was Maths-Whizz different? [Programme differentiation, Monitoring of control groups]

We expected considerable variation in the extent of differentiation between Maths-Whizz and business as usual. As relevant factors (e.g. digital literacy, assessment practices) were expected to vary significantly at teacher level, we collected teacher-level data at baseline and endpoint.

- Prior to implementation, how did teachers use formative assessment, pupil data, Edtech, and supplemental support in maths teaching?
- What activities (if any) did Maths-Whizz displace? Did it add to or displace maths learning time, including maths homework?
- To what extent did schools engage with wider Maths-Whizz resources (e.g. curriculum materials, pupil workshops) and what was the perceived impact of this?

- What was usual practice in control year groups during the intervention (e.g. supplemental support, Edtech, and formative assessment in maths)?

At design stage, we had an additional IPE RQ:

IPE RQ6: What implementation characteristics improve maths outcomes for pupils?

To understand *how* Math-Whizz affects outcomes and refine the ToC for scale-up, we wanted to explore how a small number of school-level implementation characteristics (e.g. Maths-Whizz usage, additionality of Maths-Whizz time, and home use), related to progress in a valued outcome, for example, change in maths age or maths attainment. The RQ for this analysis was:

- How do key aspects of implementation relate to a specific Maths-Whizz outcome?

Due to implementation characteristics varying significantly within each school, rather than at school level, we were not able to answer this RQ. This is discussed in the 'Analysis' section below.

IPE RQ7: What is the cost per pupil per year of Maths-Whizz? [Cost]

- What is the cost to schools per pupil per year to provide the Maths-Whizz Intelligent Tutoring programme?

Research methods

We designed an iterative IPE, with early findings influencing instrument design and sampling. The IPE lead designed the instruments and collected most of the IPE data, to ensure continuity across design, data collection, and analysis. Case study visits were undertaken by the IPE lead and one other trained qualitative researcher from NFER, who was briefed on the research aims and instruments. An overview of the IPE methods is provided in Table 8.

Table 8: IPE research methods

Set-up and baseline (November 2023 – September 2024)
IDEA workshop and set-up meetings to co-construct the TIDieR framework, understand the intervention and programme materials and refine the ToC to guide RQs. **Surveyed teachers** (intervention and control) at baseline to explore business as usual, particularly teacher efficacy and practice (e.g. formative assessment, using pupil data, Edtech, supplemental maths support). Where possible, survey questions were taken/adapted from validated teacher surveys or scales (e.g. Teaching and Learning International Survey [TALIS]).
Early implementation data (November 2024 – January 2025)
Observed two teacher training webinars (session 1), one per ESP. **Reviewed implementation plans** for all schools to understand the range of implementation approaches (e.g. extent of home learning, intended frequency of delivery, aims for using Maths-Whizz), anticipated facilitators and barriers, and goals for using Maths-Whizz. **Reviewed Maths-Whizz platform** and records for pupils (usage, progression) and parents (engagement).

Conducted a **group interview** with Maths-Whizz ESPs, exploring school support, school-level variance in early implementation, and facilitators and barriers.

Case study visits (March 2025 – April 2025)

Visited six **case study** schools (c.10% sample) chosen to allow for some variation in school characteristics and implementation approaches. The case study visits were intended to be exploratory, and from six school visits we did not expect to achieve saturation in terms of the range of practice and contextual characteristics. The in-depth qualitative exploration of case study schools was complemented by the ESP group interview, and quantitative data from the Maths-Whizz platform and surveys, which cover all schools.

Schools were sampled purposively based on FSM-levels and implementation characteristics. Schools with a high proportion of FSM-eligible pupils enabled us to explore the logic model hypotheses about differential impact for FSM pupils and the relevance of digital access at home, and to understand specific barriers and facilitators in these contexts. We considered which implementation characteristic was most important to use in the sampling frame. As the early developer interview (December 2024) and the ToC highlighted the importance of Maths-Whizz usage time as a key implementation characteristic, we chose to sample based on the average weekly usage time (time spent by pupils in tutor mode or time spent by pupils in practice mode) recorded in the Maths-Whizz platform. We used the platform data to sample schools above and below a 30-minute threshold for weekly use of Maths-Whizz.

In the trial protocol, we also intended to purposively sample schools who were implementing more integrated/long-term approaches to using Maths-Whizz (shown in blue boxes in the logic model), for example, teachers using 'topic focus' to direct the Maths-Whizz tutor to focus on particular topics relevant to the class, or using data from the Maths-Whizz platform to adapt class teaching. However, at the point of sampling, the available early data did not allow us to identify these schools. Therefore, we did not use this as a sampling criterion. However, some case study schools were using these integrated/long-term approaches. The prevalence of use of integrated/long-term approaches was explored through the endpoint teacher survey.

A stratified random sample was drawn from all schools still taking part in the intervention, based on the proportion of FSM-eligible pupils (with a threshold of 24%, based on the national average) and the average weekly pupil usage (across tutor and practice modes) recorded in Maths-Whizz between September and November (with a threshold of 30 minutes per week) (Table 9).

Table 9: Sampling frame for case study schools

		Proportion of pupils with FSMEVER6	
		Higher (>24%)	Lower (≤24%)
Average weekly pupil usage, as of 30 November 2024	Higher (≥30 mins)	2	2
	Lower (<30 mins)	1	1

While trial arm (and so intervention year groups) was not a sampling criterion because year group was not expected to moderate impact, we checked the balance of the sample by trial arm. The invited case study sample included four schools who were delivering Maths-Whizz to Years 3 and 5, and two schools who were delivering to Years 2 and 4, which allowed us to explore implementation and perceived impact for different year groups. To maximise response from sampled schools, we offered an incentive of £150 for each case study school. All six invited schools were able to accommodate a case study visit, so the achieved sample was a stratified random sample from the sampling frame in Table 9.

Case study visits included:

- Interviews with the Maths-Whizz lead (NFER's key contact for the trial and responsible for overseeing delivery of Maths-Whizz), a senior leader where different, and at least one intervention teacher. Interviews focused on their use of Maths-Whizz according to their role (training and support, use and integration, differentiation compared with business as usual, facilitators/barriers, perceived pupil engagement/outcomes, changes in cost/time). In two schools, questions for the Maths-Whizz trial lead and the senior leader were addressed by a single member of staff who performed both roles.

- A focus group in each school with Year 4 or 5 intervention pupils taught by an interviewed teacher, focusing on pupils' experience of use, engagement, and perceived outcomes. We chose Years 4 and 5 over the younger year groups because we expect older pupils to be more fluent and reflective in describing their perceptions and experiences. We aimed to include five to six pupils in each focus group, to include pupils with different characteristics while keeping the group size manageable so that all pupils were heard. Focus groups for this age group lasted a maximum of 45 minutes in order to sustain attention. Teachers were asked to select pupils in advance, including at least one pupil of higher-, medium- and lower-attainment within the class context. We asked teachers to construct a group of mixed gender, and, where possible, to include at least one FSM-eligible pupil.

Endpoint data (June 2025)

Review Maths-Whizz platform data/records (as above).

Survey school staff at endpoint. Where possible, survey questions were taken/adapted from validated teacher surveys or scales (e.g. TALIS). All questions were closed, with a mixture of single-response and multiple-response items.

- **Intervention teachers.** Training, Maths-Whizz use and integration, changes in practice, perceived pupil engagement/outcomes, moderators and mediators and costs/time, also repeating baseline questions to capture changes in efficacy/practice (e.g. formative assessment, using pupil data, Edtech, supplemental maths support).
- **Maths-Whizz trial leads and school leaders.** Onboarding/reviews, school-level implementation, changes in practice, perceived engagement/outcomes, and costs.
- **Control teachers.** Maths teaching practice during the intervention, and repeating baseline questions to explore changes and contamination.

Analysis

Qualitative Data – training observations, ESP focus groups, case study interviews, and pupil focus groups
Training observations were analysed thematically, considering fidelity/adaptations, facilitators/barriers, and teacher responsiveness. Qualitative data from the ESP interview and the six school case studies were recorded, transcribed, and analysed thematically in MAXQDA (Max Weber Qualitative Data Analysis). We developed a deductive top-level coding frame based on the RQs (e.g. fidelity and adaptations, facilitators/barriers) and coded the data from each source inductively within that frame. The inductive coding was then compared with the logic model (e.g. comparing perceived outcomes reported by teachers and pupils with the outcomes anticipated in the logic model). For the case studies, we used MAXQDA to combine the qualitative data with key implementation characteristics (e.g. extent of home use, and average pupil usage). Case-oriented thematic analysis provided a rich description of implementation for each school case (n=6).

Qualitative comparative analysis

As discussed above, we intended to categorise key school-level implementation and contextual characteristics from the teacher survey data and school implementation plans. Characteristics were initially selected based on their: i) alignment with expected moderators or other aspects of the logic model; and ii) variation across the sample of schools. Our intention was to explore this data using qualitative comparative analysis (Cilesiz and Greckhamer, 2020), a qualitative technique, which maps variation across cases (schools) and identifies the combinations of characteristics, which are consistently associated with an outcome's occurrence. With a sample size of 63 schools, we expected to be able to map four implementation and contextual characteristics, to see which combinations were associated with positive outcomes from Maths-Whizz.

The planned analysis was based on the assumption that implementation varies at school level, and implementation characteristics would be homogeneous within each school. However, on mapping the data, we found that implementation characteristics varied significantly within schools, with characteristics such as home use, the additionality of Maths-Whizz time, and achieved usage time varying primarily by class or year group. This meant that it was not feasible to undertake the qualitative comparative analysis as planned, as the data design did not allow for matching teacher and pupil data at class level, and qualitative comparative analysis is not intended for large numbers of cases (i.e. 215 classes). Instead, the initial data mapping and variation is described in the IPE results section 'Variation in implementation'.

Quantitative data: Surveys

We summarised endpoint data with descriptive statistics, including frequencies and cross-tabulations (e.g. with school-level FSM eligibility). We compared baseline and endpoint teacher data, and intervention and control teacher responses to explore changes in teacher practice, contextual changes such as other maths intervention programmes being initiated as compensation, and any self-reported contamination (e.g. use of similar maths Edtech).

Synthesis of IPE data

For each data source and participant group, we summarised the findings for each relevant RQ. We systematically compared and contrasted findings for each data source and participant group, for example, comparing perceptions of barriers and facilitators across Maths-Whizz ESPs, leaders, and teachers, and highlighting areas of congruence and difference in reporting. Where the sequence of activities allowed, we explored areas of difference directly with participants. For example, we asked case study interviewees about the reasons for any differences between the intentions recorded in school implementation plans, and the implementation in practice (based on the review of platform data).

An overview of the IPE research methods in the context of the IPE dimensions is shown in Table 10.

Table 10: IPE methods overview

Research methods	Data collection methods	Participants / data sources and achieved sample	RQs	Data analysis methods	Implementation / logic model relevance											
					Context	Fidelity	Quality of delivery	Compliance / dosage	Reach and responsiveness	Adaptation	Mediators incl. intermediary outcomes	Causal mechanisms	Moderators	Secondary outcomes	Usual practice	Cost data
IDEA workshop	TIDieR framework; logic model	The EEF, developer, NFER team	1–7	Descriptive analysis	✓	✓		✓			✓	✓	✓	✓		✓
Baseline survey	Teacher survey	Intervention and control teachers (n=269, 61% response rate)	5	Statistical analysis	✓								✓		✓	
Onboarding observations	Structured observation	Developer, schools (n=2, one per partner)	1, 3	Thematic analysis	✓	✓	✓	✓	✓	✓			✓		✓	
Review of school implementation plans	Desk review	School implementation plans (all schools who had completed implementation plans at the time of analysis, n=51)	1–4,6	Qualitative comparative analysis	✓	✓	✓	✓	✓	✓			✓		✓	
Interim Maths-Whizz platform data	Desk review	Intervention pupils (n=4,567)	1–4, 6	Statistical analysis	✓	✓	✓	✓			✓	✓	✓			
Paired interview with ESPs	Semi-structured online focus group	MW ESP (n=2)	1–4, 6	Thematic analysis	✓	✓	✓	✓		✓	✓	✓	✓	✓		
Case study visits	Interviews with school leader, trial lead, intervention class teachers. Pupil focus group Years 4 or 5	Six schools, sampled to cover a range of Maths-Whizz usage time and high/low FSM proportions. Five to six pupils per focus group (varied attainment, FSM, and gender)	1–5, 7	Thematic analysis	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Endpoint survey	Online survey	Intervention and control teachers (n=244, 65% response rate) Trial lead (n=54. 84% response rate) School leader (n=23, 47% response rate)	1–5, 7	Statistical analysis	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓
MW Platform data	MW platform data	Intervention pupils (population)	1–6	Statistical analysis		✓		✓	✓	✓	✓	✓	✓			

Costs

In line with the EEF's **cost evaluation guidance**, we conducted a cost evaluation to robustly estimate the market cost per-pupil-per-school-year of participating in Maths-Whizz, incorporating both financial and time costs. Data collection for the cost evaluation was incorporated into other planned IPE activities (below) to minimise burden to schools.

The main financial cost was expected to be the cost of the Maths-Whizz programme, which was collected directly from Maths-Whizz. As the trial inclusion criteria specified that schools should already have the technological infrastructure needed to implement Maths-Whizz, we did not expect schools to incur additional costs related to technology. However, we checked this assumption, and other financial costs, through the school leader survey and interview.

We compared the technology requirements specified for Maths-Whizz with teachers' perceptions of the requirements. We compared these requirements with Trends in International Mathematics and Science Study (TIMSS) 2023 data on digital infrastructure (available from spring 2025) to estimate the proportion of schools in England who would meet these requirements (Richardson *et al.,* 2025).

Time costs were estimated in comparison with business as usual by identifying time investments and savings in using Maths-Whizz, through the review of implementation plans, staff surveys, and staff interviews. Time costs to be considered included the time spent on training, set-up, and planning, any additional time supervising Maths-Whizz (e.g. through a homework club) and time spent monitoring, reporting, and reviewing. Time savings to be considered included reduced teacher contact time (e.g. if another member of staff is supervising Maths-Whizz use within lesson time). While we had intended to capture potential time savings relating to reduced planning time for lessons, and reduced assessment/marking time, we decided it was not feasible to capture this relatively small contribution to workload robustly and accurately without adding burden to teachers. Instead, the endpoint survey captures teachers' overall assessments of how using Maths-Whizz affected their workload.

We reviewed school implementation plans to explore the variation in the extent of home use of Maths-Whizz across schools, as we would expect home use to have different time costs and/or savings compared with school use. As all reviewed implementation plans emphasised that in-class teaching time should be used as the primary approach to Maths-Whizz, and home use was an additional option, we did not run any sensitivity analyses to estimate costs for different levels of home use.

To allow comparison with other trials, we estimated a per pupil per year cost for programme use across three years, by adding one-off costs (such as training) and annual ongoing costs (such as the financial subscription, and planning/teaching time).

Timeline

Table 11: Timeline

Dates	Activity	Staff responsible / leading
October 2023 – January 2024	• Project set-up • Complete project set-up, finalise recruitment documents and privacy notices, finalise grant agreement	Stephen Welbourne, Aarti Sahasranaman
February 2024 – May 2024	• Recruitment and pupil data collection • Finalise Data Sharing Agreement • School recruitment by Whizz Education • Recruited school data from delivery partner in three tranches – one in April and two in May • Collect pupil data for all pupils in Years 1 – 4 (will be in Years 2 – 5 in September) in participating classes	Whizz Education (recruitment), Kathryn Hurd, Jishi Jose (school and pupil data collection)

Dates	Activity	Staff responsible / leading
	• Study protocol • First draft of study protocol submitted in March 2024, revise based on feedback	Aarti Sahasranaman
June 2024 – July 2024	• Baseline impact and IPE data collection • Set-up Star Maths assessment accounts for schools and pupils • Coordinate baseline assessments (Star Maths assessment, Maths and Me survey) in schools • Baseline teacher survey	Kathryn Hurd, Jishi Jose, Katherine Aston (teacher survey)
	• Randomisation • Randomisation of schools in July 2024	Gemma Schwendel
August 2024 – July 2025	• Maths-Whizz training, delivery, and data sharing (by Whizz Education) • Teacher training and onboarding in August 2024 – September 2024 • Maths-Whizz delivery in schools from September 2024 – July 2025 • Maths-Whizz share periodic platform usage reports with NFER throughout delivery period	Whizz Education
	• Statistical Analysis Plan • First draft of Statistical Analysis Plan submitted in September 2024, edits based on the EEF, peer reviewer and delivery partners' feedback, final Statistical Analysis Plan published in January 2025	Gemma Schwendel
	• IPE • Observation of teacher training and onboarding in August 2024 – September 2024 • IPE instrument development in December 2024 – January 2025 • IPE activities (case study visits, ESP focus groups) in January 2025 – March 2025	Katherine Aston
	• NPD application • Submit NPD application in October 2024	Gemma Schwendel
	• Endpoint assessments • Endpoint assessments by NFER test administrators in June 2025 – July 2025	Kathryn Hurd, Jishi Jose
July 2025 – August 2025	• Data cleaning	Kathryn Hurd
September 2025 – November 2025	• Analysis • Access matched pupil-level dataset on the Office for National Statistics Secure Research Service, complete primary and additional analyses	Gemma Schwendel, Aarti Sahasranaman
December 2025	• Reporting • Submit first draft of report	Aarti Sahasranaman, Katherine Aston, Stephen Welbourne
May 2025 – June 2026	• Final report published	Aarti Sahasranaman, Katherine Aston, Stephen Welbourne, EEF
August 2026	• Data archiving • Submit data to the EEF archive	Gemma Schwendel

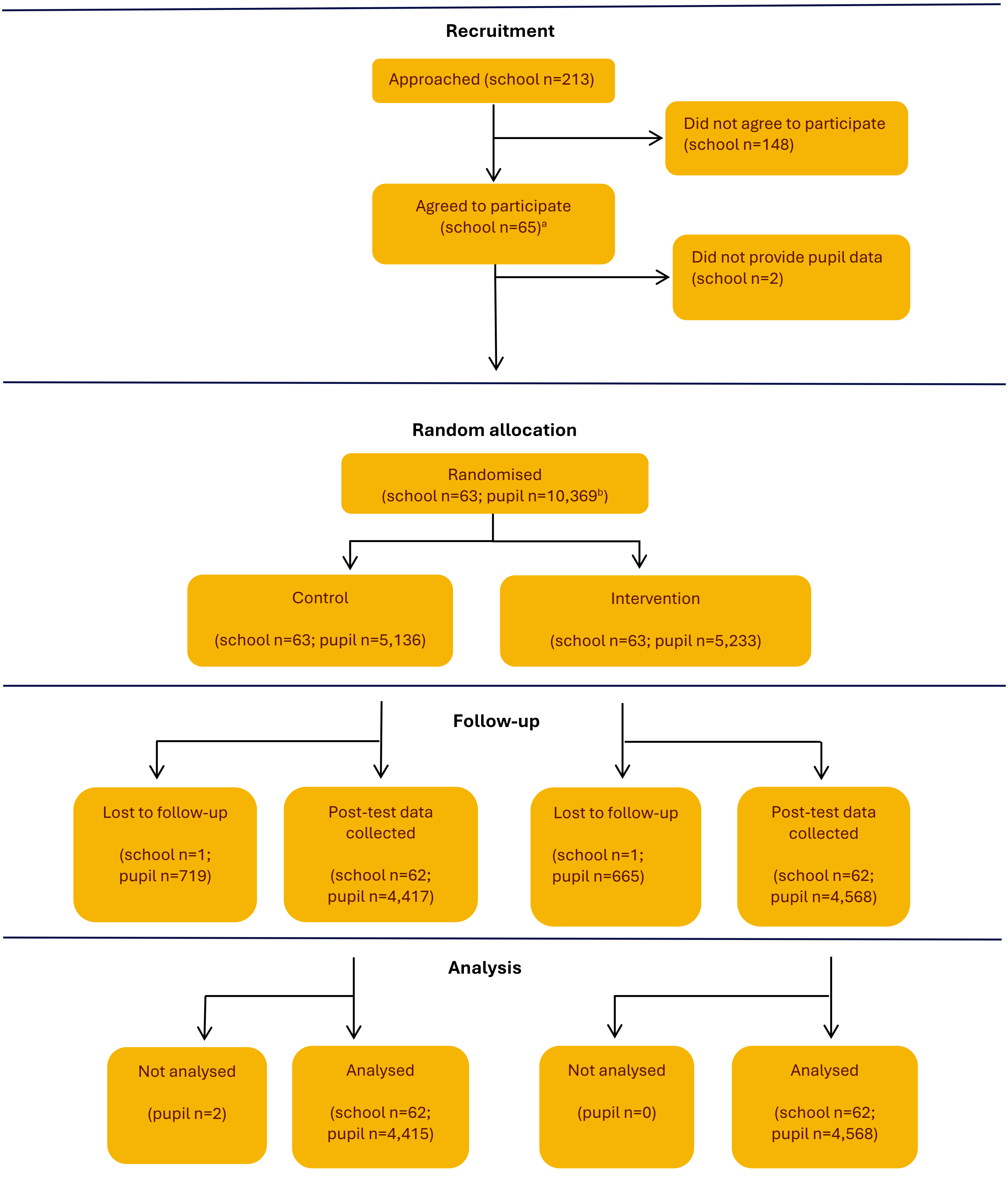
Impact evaluation results

Participant flow including losses and exclusions

The participant flow diagram for this evaluation is shown below (Figure 2). In the recruitment phase, 95 schools submitted an EOI. Of these, 68 schools signed the MoU signalling their commitment to participate in the trial. Two of these 68 schools dropped out when NFER approached them for pupil data collection. One school dropped out because they were found to be ineligible and the other had changes in staff capacity making their continued participation in the trial difficult. Of the remaining 66 schools that provided pupil data, six schools were either infant or junior schools and at randomisation, each of these three pairs of infant and junior schools were considered as a single cluster or school. On this basis, a total of 63 schools/clusters were randomised.

A total of 444 classes from across Years 2, 3, 4, and 5 participated in this trial. Of the 10,369 pupils who were randomised, 5,136 were allocated to the control and 5,233 to the intervention group. Around 1,384 pupils did not provide endpoint Star Maths Unified Scaled Scores (control N=719, intervention N=665). A key contributor to pupil attrition was the withdrawal of one school from the trial. Other reasons included pupils being absent on the day of the endpoint assessment, pupils leaving the school, pupils withdrawn from the assessment because of additional needs that made it difficult to access the Star Maths assessment, and internet connectivity issues that caused some assessments to be paused midway. An additional two pupils who provided endpoint data were excluded from the final analysis sample because they were missing class information. The final primary analysis sample, therefore, consisted of 8,983 pupils (control N=4,415, intervention N=4,568).

Figure 2: Participant flow diagram (two arms)



[a] A total of 68 schools signed the MoU indicating their agreement to participate in the trial. This included three infant schools and three junior schools. For the purpose of the trial, each of the three pairs of infant and junior schools were considered a single cluster or school. This reduced the number of individual clusters or schools to 65, as shown in the figure.

[b] While 10,383 pupils were initially present at randomisation, 14 were subsequently withdrawn from data processing, and were not invited for endpoint testing.

The MDES at protocol, randomisation, and analysis stages is shown in Table 12. A 10% attrition rate was predicted between recruitment and endpoint testing: this quantity has been deducted from the number of schools in the protocol and analysis stages. A 15% pupil-level attrition rate was also assumed for the protocol and randomisation stages. At the analysis stage, the parameters in the MDES calculations were more favourable than anticipated in the protocol: pupil attrition was lower than 15% and pre-/post-correlation was higher. As a result, the MDES for all pupils reduced from 0.11 at protocol stage to 0.061 at analysis stage. The proportion of FSM pupils (25.6%) was slightly higher than those assumed at the protocol and randomisation stages (24.6%) leading to a larger FSM sample. Alongside this, the key MDES drivers moved in ways that mostly improved precision and reduced MDES. The pupil-level pre-/post-correlation increased materially, directly lowering MDES. Second, the assumptions about ICC were not borne out. The introduction of class-level ICC at the analysis stage (which was not included in earlier MDES calculations) redirected some of the total variance away from the school level. This reduction in school-level ICC thus decreases the MDES.

While the school-level pre-/post-correlation dropped from 0.263 to 0 in the analysis—which could modestly reduce the benefit of baseline adjustment at that level—the combination of stronger pupil/year group correlations and lower school ICC dominates.

Table 12: MDES at different stages

		Protocol		Randomisation		Analysis	
		Overall	FSM	Overall	FSM	Overall	FSM
MDES		0.11	0.13	0.09	0.12	0.06	0.11
Pre-test/post-test correlations[a]	Level 1 (pupil)	0.38	0.38	0.38	0.38	0.53	0.54
	Level 2 (class)						
	Level 3 (year group)					0.95	0.94
	Level 4 (school – KS1)	0.26	0.26	0.26	0.26	0	0
ICCs	Level 2 (class)					0.03	0.03
	Level 3 (year group)	0.073	0.073	0.073	0.073	0.066	0.066
	Level 4 (school)	0.036	0.036	0.036	0.036	0.02	0.01
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?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average pupils per class		26.7		23.4		21.5	6.25
Average classes per year group		1.4	1.4	1.7	1.4	1.68	1.60
Proportion of pupils eligible for FSM within last six years			24.6%		24.6%		25.6%
Number of pupils	Intervention	4,336	1,067	5,252[b]	1,292	4,568	1,176
	Control	4,336	1,067	5,131[b]	1,262	4,415	1,125
	Total:	8,672	2,133	10,383[b]	2,554	8,983	2,301
Number of classes	Intervention	162	162	223	223	215	190
	Control	162	162	221	221	198	187

	Total:	325	325	444	444	405[c]	368
Number of schools	Total:	58	58	63	63	62	62

[a] R2 for the school level is set to zero as the null model shows no variance at the school level.
[b] The post-randomisation sample size calculations were performed immediately post-randomisation, before any withdrawal of pupils from data processing.
[c] The number of classes at the analysis stage exceeds that reported elsewhere due to a few mixed year group classes. Because randomisation occurred at the year group level, these classes are counted in both groups.

Attrition

Overall, the pupil attrition rate between randomisation and primary analysis was moderate at 13.4%, with a higher attrition rate seen among control pupils (14.0%) compared to intervention pupils (12.7%) (Table 13). One school withdrew from the study. Due to the interleaved nature of the design, this resulted in attrition across both intervention and control groups (235 pupils altogether; 121 in the intervention group and 114 in the control group). Other reasons for missing pupil endpoint data were pupil absence on test day, pupils leaving the school, withdrawal of pupils from testing by the teacher because of additional needs that made accessing the Star Maths assessment untenable and internet connectivity issues, which caused some tests to be paused midway. This moderate level of attrition was expected: a 15% pupil-level attrition rate was assumed in the MDES calculations given the ongoing challenges with school absence, even in primary schools.

Table 13: Pupil-level attrition from the trial (primary outcome)

		Control	Intervention	Total
Number of pupils	Randomised	5,136	5,233	10,369
	Analysed	4,415	4,568	8,983
Pupil attrition (from randomisation to analysis)	Number	721	665	1,386
	Percentage	14.0%	12.7%	13.4%

Pupil and school characteristics

The balance at baseline for the two groups has been analysed as randomised against a number of characteristics at the school level (Table 14): and at the pupil level (Table 15).

Due to the interleaved nature of the design (i.e. with schools present in both intervention and control groups), school-level characteristics are presented for the sample overall and compared to the national figures for the following characteristics:

- proportion of FSM-eligible pupils within the school;
- if the school is rural or urban;
- type of school;
- region; and
- the latest Ofsted rating.

The following pupil-level characteristics at randomisation are presented separately for intervention and control groups:

- gender;
- year group;
- FSM eligibility; and
- prior attainment level.

Analysis of pupil-level characteristics shows negligible differences between intervention and control groups across all characteristics (Table 14). This is supported by the baseline effect size (0.01; 95% CI: −0.28 to 0.30), which is very small and spans zero. None of these differences are statistically significant, so weighting is not warranted in subsequent analyses. The distribution of characteristics among analysed pupils remained broadly consistent with baseline (Table 15). One school withdrew during the intervention; it had been randomised with Years 2 and 4 in the intervention group and Years 3 and 5 in the control group. This withdrawal slightly altered year group proportions but again not enough to justify weighting.

Table 14: Baseline characteristics of trial schools compared with all eligible primary schools in England

School level (categorical)	Sample		National level	
	N	%	N	%
Urban/rural				
Rural	7	11.0%	4854	29.0%
Urban or Other	56	89.0%	11875	71.0%
Region				
East Midlands	9	14.3%	1653	9.9%
London	10	15.9%	1777	10.6%
North East	6	9.5%	859	5.1%
North West	12	19.0%	2435	14.6%
South East	13	20.6%	2594	15.5%
West Midlands	7	11.1%	1777	10.6%
Other[a]	6	9.5%	5,634	33.7%
FSM quintile				
Lowest 20%	11	17.5%	3364	20.1%
Second Lowest 20%	8	12.7%	3348	20.0%
Middle 20%	16	25.4%	3344	20.0%
Second Highest 20%	13	20.6%	3326	19.9%
Highest 20%	17	27.0%	3324	19.9%
Missing	0	0.0%	23	0.1%
Type of school				
Academies	32	50.8%	7475	44.7%
Free Schools	4	6.3%	285	1.7%
LA Maintained	27	42.9%	8969	53.6%
Ofsted rating				
Outstanding	8	13.1%	1565	9.4%
Good	48	78.7%	12871	76.9%
Requires Improvement / Inadequate / Not judged / Missing	7	11.1%	2,293	13.7%

[a] *Three regions have been grouped into 'Other' to avoid disclosure.*

Table 15: Baseline characteristics of trial pupils

Pupil level (categorical)		Intervention		Control		
		n/N (missing)	Count (%)	n/N (missing)	Count (%)	
FSM eligibility						
No		3,883/5,233 (0)	74%	3,835/5,136 (0)	75%	
Yes		1,350/5,233 (0)	26%	1,301/5,136 (0)	25%	
Gender						
Female		2,519/5,233 (0)	48%	2,531/5,136 (0)	49%	
Male		2,714/5,233 (0)	52%	2,605/5,136 (0)	51%	
Year group						
Year 2		1,227/5,233 (0)	23%	1,273/5,136 (0)	25%	
Year 3		1,340/5,233 (0)	26%	1,246/5,136 (0)	24%	
Year 4		1,298/5,233 (0)	25%	1,333/5,136 (0)	26%	
Year 5		1,368/5,233 (0)	26%	1,284/5,136 (0)	25%	
Prior attainment						
Low prior attainment		838/5,233 (0)	16%	819/5,136 (0)	16%	
Not low prior attainment		4,395/5,233 (0)	84%	4,317/5,136 (0)	84%	
Pupil level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size (95% CI)
Baseline Star Maths Unified Scaled Score	Not available at the national level	5,233/5,233 (0)	905.9 (85.02)	5136/5136 (0)	905.55 (85.10)	-0.01 (0.08,0.06)

Outcomes and analysis

Primary analysis

The mean Star Maths Unified Scaled Score for pupils was 957.7 overall, with a six-point difference between the means for the intervention and control groups (960.8 and 954.6, respectively). Pupils who received access to Maths-Whizz made, on average, an additional month's progress in maths compared to control pupils. This corresponds to an effect size of 0.08 (95% CI: 0.02, 0.14, see Table 16), with a p-value of 0.01. This is the best estimate of the intervention's impact but the possible impact of this programme also includes no progress and positive effects of up to two months' additional progress.

As shown in Figure 3 below, the distribution of the endpoint Star Maths Unified Scaled Score is broadly normally distributed (Gaussian) with little difference in the distribution between the intervention and control groups, similar to that seen at baseline (see Appendix D). Although both groups showed a general improvement in the scores between baseline and endpoint, the mean improvement is slightly higher among pupils in the intervention group compared to the control group (see Table 16 below) This holds across all year groups, though is particularly notable for Years 2 and 3 (Table 18).

Table 16: Primary outcome analysis results

Outcome	Unadjusted means				Effect size			
	Intervention group		Control group					
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	ICC (full)	Hedges' g (95% CI)	P-value
Star Maths Unified Scaled Score	4,568 (665)	960.80 (792.36, 1129.26)	4,415 (665)	954.60 (783.58, 1125.57)	8,983 (4,415)	School: 0.02 Year: 0.07 Class: 0.03 Total: 0.11	0.08 (0.02, 0.14)	0.010

The effect size presented in Table 16 is calculated using the difference between intervention and control group means ($\acute{o}_i - \acute{o}_c$), based on Unified Scaled Scores, as the numerator. The denominator is the pooled SD calculated across eight groups (intervention and control within each year group), following the approach described earlier in the section 'Estimation of effect sizes'.

To aid clarity and provide a robustness check, Table 17 presents effect sizes derived using several alternative calculation methods. When the pooled SD is computed across only two groups (intervention vs control), the resulting effect size is smaller. This is because this approach captures additional age-related variation (across four of the year groups) when Unified Scaled Scores are used. In contrast, approaches that control for age—either by extending the pooled SD to eight groups and/or by using age-standardised scores when calculating the numerator—produce larger and highly consistent effect sizes.

Because the unadjusted mean difference and its 95% CI are calculated using Unified Scaled Scores, the effect size reported in Table 16 is the one based on Unified Scaled Scores with the pooled SD calculated across eight groups. This method is used consistently throughout the report wherever effect sizes are presented.

Table 17: Additional effect size calculations

Effect size calculation	Effect size (95% CI)	P-value
Numerator: Difference based on Unified Scaled Scores; Denominator: Pooled SD over two groups (intervention and control)	0.06 (0.01,0.11)	0.011
Numerator: Difference based on Unified Scaled Scores; Denominator: Pooled SD over eight groups (intervention and control within year group)	0.08 (0.02, 0.14)	0.011
Numerator: Difference based on age-standardised scores Denominator: Pooled SD over two groups (intervention and control)	0.08 (0.02, 0.13)	0.011
Numerator: Difference based on age-standardised scores Denominator: Pooled SD over eight groups (intervention and control within year group)	0.08 (0.02, 0.13)	0.011

Figure 3: Histogram of endpoint Star Maths Unified Scaled Scores for control and intervention pupils

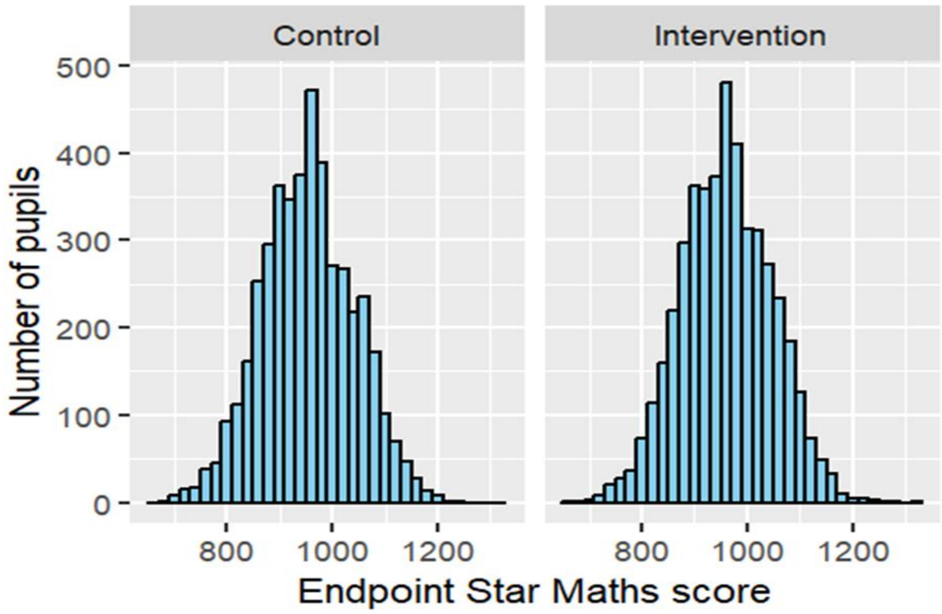


Table 18: Mean difference (standard error) in Star Maths Unified Scaled Scores between baseline and endpoint, split by control and intervention group and year group

Year group	Intervention	Control
Year 2	56.69 (1.66)	46.41 (1.76)
Year 3	49.82 (1.41)	41.94 (1.56)
Year 4	48.70 (1.40)	41.95 (1.39)
Year 5	53.28 (1.51)	51.66 (1.49)

Secondary analysis

Subgroup analyses (RQ2 to RQ5)

The following subgroup analyses were conducted to examine whether the impact of Maths-Whizz varied by:

1. year group (RQ2);
2. FSM eligibility (RQ3);
3. gender (RQ4); and
4. prior attainment (RQ5).

The results of these analyses are shown in Table 19. The FSM subgroup analysis shows a positive effect for FSM-eligible pupils (0.14, 95% CI: 0.06 to 0.22). This effect is larger than for the primary research sample equating to two months' additional progress. The year group analysis indicated small positive effects for all year groups, with the smallest for Year 5 (0.02, 95% CI: -0.08 to 0.12) and the largest effect for Year 3 (0.13, 95% CI: 0.03 to 0.24), equivalent to two months' progress, suggesting that Maths-Whizz may be most effective for this year group. Year 3 was also the only year group where the 95% CI did not include zero. However, we would caution against reading too much into the differences between year groups as the CIs are all overlapping and there was no evidence to support an interaction between year group and intervention. It should also be noted that the year group analysis had reduced statistical power compared with the primary analysis arising from smaller sample sizes.

The school-level ICCs reported in Table 19 vary across year groups, indicating differing degrees of outcome clustering between schools. The ICC for Year 2 pupils (0.17) is substantially higher than for other year groups, implying that a larger proportion of variance is attributable to between-school differences rather than individual pupils. This stronger clustering inflates the standard error of the model-based estimate for Year 2 pupils, reducing the precision of the corresponding effect size. Consequently, although the raw difference in mean change between intervention and control groups is relatively large for Year 2 (Table 18), the effect size derived from the model is smaller and statistically non-significant. In contrast, Year 3 exhibits a much lower ICC (0.05), indicating weaker clustering, greater statistical precision, and a larger, statistically significant effect size. The Year 2 results should therefore, be interpreted as reflecting greater uncertainty due to school-level clustering, rather than evidence that the intervention was ineffective.

The effect sizes for all other subgroups (boys, girls, pupils with low prior attainment, and those with higher prior attainment) were all small, positive, and with CIs that were above zero. The effect size was higher for boys than girls (0.10 and 0.07, respectively), and higher among pupils with low prior attainment compared to those with higher prior attainment (0.16 and 0.08, respectively).

The differential impact of the intervention for pupils in each subgroup (e.g. FSM pupils compared with non-FSM pupils) is additionally shown in Table 20, columns 3 and 4. With the exception of FSM eligibility, all effect sizes for the interaction between intervention and subgroup (column 3) are close to zero with 95% CIs that span zero, suggesting that the intervention is similarly effective for these subgroups, and any observed differences are likely due to sampling variability rather than a systematic pattern. The 95% CI for the interaction term for FSM eligibility with grouping sits above zero, indicating that the effect of Maths-Whizz on maths attainment differs according to FSM eligibility, with a larger effect seen amongst FSM-eligible pupils.

As required by the EEF statistical analysis guidance (EEF, 2022), we have also estimated the effect size within each subgroup by combining the appropriate terms from the interaction model (Table 20, columns 5 and 6). This alternative approach to obtaining the subgroup effect sizes produced estimates that were largely similar to the effect sizes obtained from the interaction model in Table 19.

Table 19: The main analysis model (Star Maths Unified Scaled Score) restricted to particular subgroups of pupils

Analysis	Unadjusted means					Effect size			
		Intervention group		Control group					
	Subgroup	N (missing)	Mean (95% CI)	N (missing)	Mean (95% CI)	Total N (intervention; control)	ICC (full)	Hedges' g (95% CI)	P-value
Year group	Year 2	1,060 (167)	886.91 (766.70, 1007.1)	1,089 (184)	885.31 (760.97, 1009.66)	2,149 (1,060, 1,089)	School: 0.17 Class: 0.02 Total: 0.194	0.07 (-0.11, 0.26)	0.44
	Year 3	1,173 (167)	940.48 (814.02, 1066.93)	1,038 (208)	928.48 (805.39, 1051.58)	2,211 (1,173, 1,038)	School: 0.05 Class: 0.01 Total: 0.06	0.13 (0.03, 0.234)	0.01
	Year 4	1,108 (190)	979.39 (845.51, 1113.27)	1,165 (168)	973.64 (832.31, 1114.96)	2,273 (1,108, 1,165)	School: 0.04 Class: 0.04 Total: 0.08	0.08 (-0.01, 0.18)	0.09
	Year 5	1,227 (141)	1027.32 (872.78, 1181.86)	1,123 (161)	1026.08 (871.50, 1180.66)	2,350 (1,227, 1,123)	School: 0.06 Class: 0.02 Total: 0.08	0.02 (-0.08, 0.12)	0.67
FSM	Eligible for FSM	1,176	940.2 (768.50,1111.90)	1,125	930.4 (763.02,1097.78)	2,301 (1,176, 1,125)	School: 0.01 Year: 0.07 Class: 0.00 Total: 0.08	0.14 (0.06,0.22)	0.01
Gender	Boys	2,362 (352)	969.42 (794.87, 1106.93)	2,249 (356)	962.57 (783.54, 1141.60)	4,611 (2,362, 2,249)	School: 0.03 Year: 0.06 Class: 0.02 Total: 0.10	0.10 (0.03, 0.16)	<0.01
	Girls	2,206 (313)	951.60 (791.86, 1111.34)	2,166 (365)	946.27 (785.62, 1106.93)	4,372 (2,206, 2,166)	School: 0.02 Year: 0.07 Class: 0.03 Total: 0.12	0.07 (0.00,0.14)	0.05
Prior attainment	Low prior attainment	663 (175)	875.61 (742.62,1008.60)	644 (175)	867.97 (733.55, 1002.39)	1,307 (663, 644)	School: 0.04 Year: 0.03 Class: 0.00 Total: 0.07	0.16 (0.05, 0.28)	0.01
	Not low prior attainment	3,905 (490)	975.28 (818.26, 1132.29)	3,771 (546)	969.36 (810.02, 1128.71)	7,676 (3,905, 3,771)	School: 0.02 Year: 0.08 Class: 0.02 Total: 0.12	0.08 (0.01, 0.15)	0.03

Table 20: The main analysis model (Star Maths Unified Scaled Score) with interaction term for subgroup analysis

Subgroup (1)	Total n (missing) (2)	Subgroup x intervention (differential impact of Maths-Whizz for subgroup pupils)		Intervention + Subgroup x intervention (impact in the subgroup)	
		Effect size (95% CI) (3)	P-value (4)	Effect size (95% CI) (5)	P-value (6)
Year group: Year 2	8,983 (1,386)	0.01 (-0.13,0.14)	0.93	0.08 (-0.07, 0.23)	0.28
Year group: Year 3	8,983 (1,386)	0.04 (-0.1,0.18)	0.93	0.12 (-0.04, 0.27)	0.16
Year group: Year 4	8,983 (1,386)	0.00 (-0.14,0.15)	0.99	0.08 (-0.08, 0.24)	0.35
Year group: Year 5	8,983 (1,386)	-0.06 (-0.18,0.07)	0.38	0.02 (-0.12, 0.16)	0.76
FSM	8,983 (1,386)	0.08 (0.01,0.14)	0.03	0.13 (0.06,0.21)	<0.01
Gender	8,983 (1,386)	0.03 (-0.04,0.08)	0.37	0.09 (0, 0.17)	0.04
Prior attainment	8,983 (1,386)	0.04 (-0.05,0.11)	0.34	0.11 (0, 0.22)	0.04

Mathematical self-perceptions and enjoyment of mathematics (RQ6)

Overall, the mean scores for the self-perception and enjoyment subscales of the Maths and Me survey were 28.6 and 36.0, respectively. There were some small differences in the means between intervention and control groups (self-perception: 28.9 and 28.2; enjoyment: 36.4 and 35.6, respectively, see Table 21). For both subscales, the Hedges' g effect size was small (0.05 and 0.06 for self-perception and enjoyment, respectively), with the CI just crossing zero in both cases, suggesting that using Maths-Whizz probably results in a small positive effect on pupils' self-perceptions and enjoyment of maths.

Table 21: Secondary analysis: Maths and Me survey self-perception and enjoyment subscales

Outcome	Unadjusted means				Effect size		
	Intervention group		Control group				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	Hedges' g (95% CI)	P-value
Self-perception	3,345 (534)	28.85 (13.92, 43.78)	3,118 (532)	28.24 (13.32, 43.15)	6,463 (3,118; 3,345)	0.05 (-0.00, 0.11)	0.07
Enjoyment		36.42 (15.20, 57.64)		35.56 (14.23, 56.90)		0.06 (-0.01, 0.12)	0.08

Additional analyses and robustness checks

Dosage

The relationship between programme dosage and pupils' maths attainment was examined using usage data from the Maths-Whizz platform, specifically time spent in tutor mode and practice mode (Table 22) over the complete duration of the intervention, including school holidays. We included usage during the school holidays in the dosage analysis as some pupils were given home access and could potentially have accessed Maths-Whizz during holidays. On average, pupils engaged for just under 25 minutes per week in tutor mode and around eight minutes per week in practice mode. While average time spent in tutor mode varied slightly across year groups (Table 22), these differences are small and fall within ranges that broadly overlap. In contrast, the pattern for practice mode is more distinct: younger pupils tend to spend noticeably more time in practice mode than older pupils, with each year group showing a range that does not intersect with the others.

Table 22: Average weekly time spent in tutor and practice modes on Maths-Whizz

	N	Mean tutor mode (95% CI)	Mean practice mode (95% CI)
Overall	**4,567**	**24.59 (24.23, 24.94)**	**7.62 (7.47, 7.78)**
Year 2	1,060	24.46 (23.78, 25.13)	9.37 (9.01, 9.74)
Year 3	1,172	25.35 (24.72, 25.98)	7.98 (7.67, 8.28)
Year 4	1,108	24.39 (23.49, 25.28)	6.97 (6.69, 7.25)
Year 5	1,227	24.15 (23.54, 24.76)	6.36 (6.12, 6.61)

Although both modes were positively associated with maths attainment, with higher usage linked to greater improvement, time spent in practice mode had a greater impact on maths attainment (0.73, 95% CI: 0.40 to 1.05). Specifically, an additional 15 minutes per week was associated, on average, with an 11-point increase in scaled score for practice mode, compared with a 3-point increase for tutor mode. The greater impact of time spent on practice mode could reflect the requirement that pupils could use the practice mode only after they achieved three progressions on the tutor mode. Therefore, time spent on practice mode may reflect high attainment (rather than practice time mode *per se*) as the pathway through which the impact on attainment is produced. The best guide to the effect of different levels of dosage is therefore time spent in tutor mode. For pupils who achieved the recommended mean weekly use of 45 minutes in tutor mode, this would result in a 9-point increase in scaled score, equating to two additional months' progress.

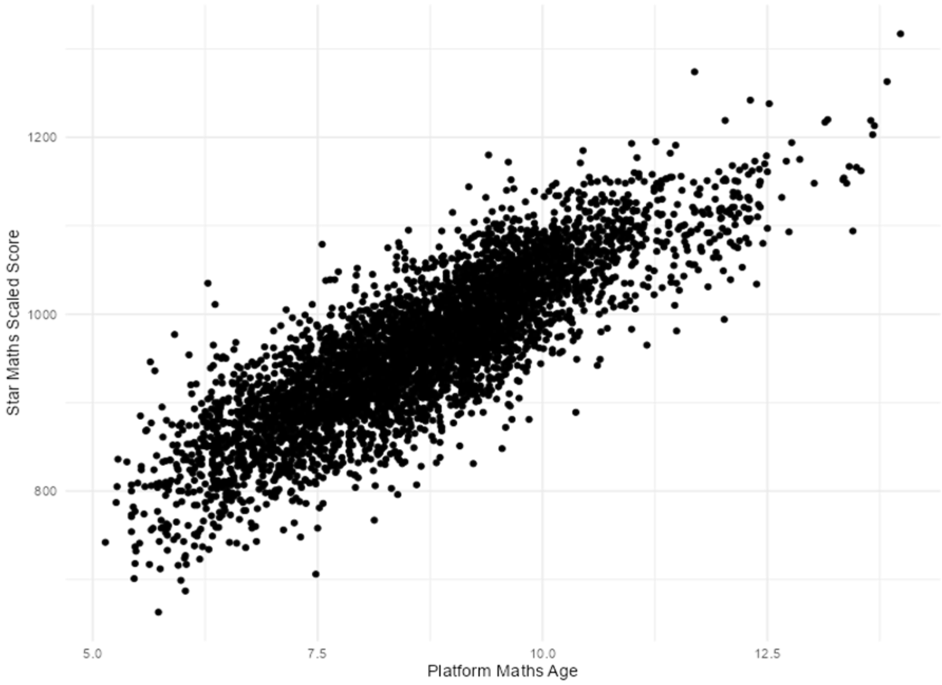
Table 23: Dosage analysis

Variable	n (missing)	Mean time spent (minutes)	Estimate (95% CI)	P-value
Tutor Use	4,567 (10)	24.6	0.20 (0.054, 0.34)	0.01
Practice Mode		7.6	0.73 (0.40, 1.05)	<0.0001

Correlation between Maths-Whizz maths age and maths attainment

A correlation analysis was conducted to examine the relationship between endpoint maths age (as determined by the Maths-Whizz tutor) and maths outcomes measured by endpoint Star Maths Unified Scaled Scores (Figure 4). The correlation coefficient was strong ($r = 0.844$, $p<0.0001$), indicating a robust positive association between maths age and externally assessed maths attainment. This provides strong evidence supporting the validity of the maths age metric generated by the app.

Figure 4: Scatterplot between Star Maths Unified Scaled Score and maths age



Analysis in the presence of non-compliance

An analysis was completed on the outcomes of pupils who were in classes considered to be compliant. A class was deemed to have been compliant with the intervention if:

- Pupils in the class spent at least 30 minutes tutor usage per week, on average, in the total period considered (term-time only/excluding the holidays).

  OR

- Pupils in the class attempted three tutor lessons per week (pass, static, or fail), on average, in the total period considered (term-time only/excluding the holidays).

Of the 215 classes within the intervention group, 195 (91%) classes across 61 schools were compliant, meeting at least one of the two criteria (Table 24). The majority of classes (120) were compliant only according to the number of tutor lessons delivered criteria, with pupils attempting at least three tutor lessons per week on average.

Table 24: Compliance at the class level

		Tutor usage	
		Compliant	Not compliant
Tutor lessons	Compliant	75	120
	Not compliant	0	20

Around 95% of intervention pupils (4,183) were in classes deemed to have been compliant. Table 25 shows the impact of Maths-Whizz on mathematical attainment when pupils were in classes that met at least one of the compliance criteria. There is a positive effect size (0.11) for these pupils, with the whole 95% CI lying above zero, suggesting that Maths-Whizz has a beneficial effect (equivalent to two additional months' progress) on mathematical attainment among pupils where Maths-Whizz is used in this way. This is the best estimate of the impact of compliance to the intervention, but the possible impacts also include positive effects of one additional month's progress.

Table 25: Analysis in the presence of non-compliance

Model stage	Total n (intervention, control)	Predictor	Standardised effect size (95% CI)	P-value
Stage 1: Compliance indicator regressed on intervention status	8,983 (4,415, 4,568)	Intervention status	3.73 (3.69, 3.77)	< 0.0001
Stage 2: Star Maths Unified Scaled Score regressed on compliance indicator		Compliance indicator	0.11 (0.08, 0.15)	< 0.0001

Missing data analysis

A missing data analysis was run to determine whether there were any patterns in missingness of the Star Maths endpoint assessment, and whether this needed to be addressed via multiple imputation. Overall, 1,384 pupils (13%) were missing an endpoint Star Maths Unified Scaled Score, with a slightly higher proportion in the intervention group compared to the control group (14% vs 13%). Table 26 below presents the results of a multilevel logistic regression conducted to assess whether any included covariates predicted missing endpoint assessment scores.

Table 26: Multilevel logistic regression where the outcome is whether a pupil is missing their endpoint Star Maths Unified Scaled Score

Level	Variable[a]	N (%) covariate missing	Odds ratio (OR) (95% CI)	P-value
Pupil level	FSM status: Eligible	0 (0%)	1.32 (1.13,1.54)	<0.001
	Gender: Male	0 (0%)	1.04 (0.91, 1.19)	0.53
	Year group: Year 3	0 (0%)	1.13 (0.90,1.42)	0.29
	Year group: Year 4	0 (0%)	1.09 (0.84,1.41)	0.52
	Year group: Year 5	0 (0%)	0.96 (0.71,1.28)	0.77
	Absence rate	206 (2%)	1.05 (1.04, 1.06)	< 0.001
	Low prior attainment	0 (0%)	1.28 (1.01,1.61)	0.04
School	Urban/rural: Urban	0 (0%)	1.17 (0.56,2.43)	0.67
	FSM quintile: Second lowest 20%	0 (0%)	1.42 (0.73,2.73)	0.30
	FSM quintile: Middle 20%	0 (0%)	0.87 (0.45,1.69)	0.68
	FSM quintile: 2nd Highest 20%	0 (0%)	1.07 (0.54,2.10)	0.85
	FSM quintile: Highest 20%	0 (0%)	1.30 (0.64,2.63)	0.47
	Ofsted rating: Outstanding	0 (0%)	0.80 (0.40,1.61)	0.53
	Ofsted rating: Requires Improvement	0 (0%)	0.76 (0.32,1.79)	0.53
	Ofsted rating: Not Judged	0 (0%)	0.93 (0.26,3.29)	0.91

[a] Baseline categories for factor variables are FSM eligible: Ineligible; Gender: Female; Year group: Year 2; Urban/rural: Rural; FSM quintile: Lowest 20%; Ofsted rating: Good.

Findings indicate that none of the school-level covariates were associated with an increased likelihood of missing data. At the pupil level, FSM status, low prior attainment, and absence rate were associated with missing endpoint assessment scores. The 95% CIs for the ORs indicated that pupils eligible for FSM, with low prior attainment and/or with higher absence rates were more likely to have missing outcome data. To assess whether these predictors of missingness also confounded the estimated intervention effect, we re-ran the primary outcome model including all three variables as covariates on a complete-case basis. Although all pupils had a low prior attainment flag and FSM status, eight cases had missing absence rate and were subsequently excluded from the model. As shown in Table 27, including these covariates in the primary outcome model did not meaningfully change the estimated treatment effect. This stability suggests that the primary finding is robust to adjustment for plausible missingness mechanisms, and that any bias arising from missing data is negligible. The EEF statistical guidance (EEF, 2022) recommends the use of multiple imputation when variables associated with missingness are also expected to confound the treatment effect (in this case, absence rate). In this evaluation, however, the number of missing records for these variables was very small, and the sensitivity analysis indicated that they do not substantively confound the treatment–outcome relationship. Given the lack of evidence for meaningful confounding and the negligible impact of adjusting for predictors of missingness, complete-case analysis was considered adequate. Multiple imputation was therefore, not pursued.

Table 27: Primary outcome sensitivity analysis results

Outcome	Unadjusted means				Effect size			
	Intervention group		Control group					
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	Total n (intervention; control)	ICC (full)	Hedges' g (95% CI)	P-value
Star Maths Unified Scaled Score	4,566 (667)	960.82 (792.34, 1129.29)	4,409 (727)	954.57 (783.58, 1125.57)	8,975 (4,409, 4,562)	School: 0.02 Year: 0.07 Class: 0.04 Total: 0.11	0.08 (0.02, 0.14)	0.01

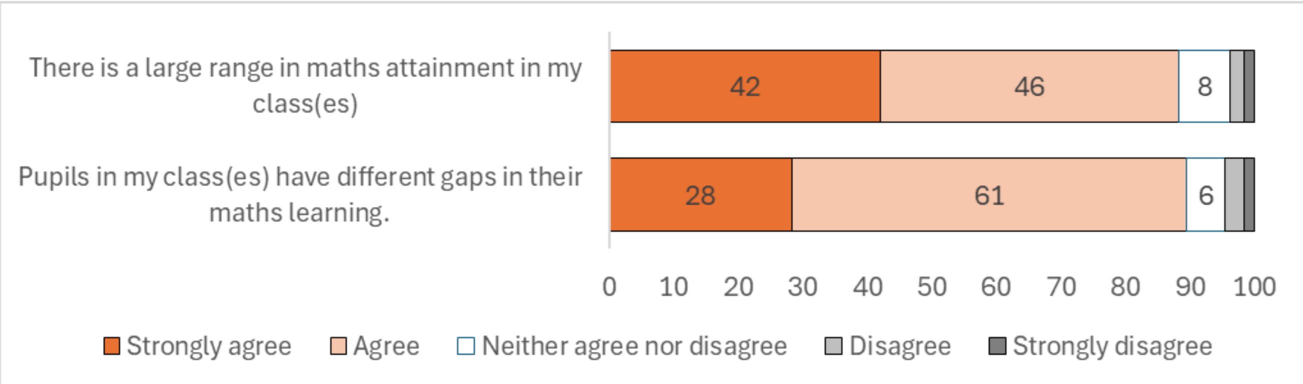
Implementation and Process Evalaution results

This section outlines the context of the trial, including teachers' experience of the problems, which Maths-Whizz aims to address, and their teaching context and experience, and then describes the usual practice with which Maths-Whizz is being compared in this trial. The process of Maths-Whizz implementation encourages schools to make contextualised decisions about how and when to use Maths-Whizz, so this variation is described, as well as key aspects of fidelity, which all schools are asked to adhere to. It explores how teachers and pupils responded to Maths-Whizz, and the perceived impact for pupils and teachers. This evidence is compared with the logic model, including exploring perceptions of key moderators and pathways to impact. Throughout this section, it is important to note that endpoint surveys tend to be biased towards schools/staff with higher engagement and/or more positive views, as withdrawn or disengaged schools and staff are less likely to respond.

Context

Maths-Whizz aims to address two key problems (as described in the logic model): the four-year knowledge gap in a typical primary maths class (Whizz Education, 2021), and the individual nature of pupils' learning gaps. Based on baseline teacher survey data, which reflects teachers' experience of their previous classes (2023/2024) before their trial classes (2024/2025), these problems were highly relevant to almost all trial teachers, with 88% agreeing that there is a large range in maths attainment in their class, and 89% agreeing that their pupils have different gaps in their maths learning (Figure 5).

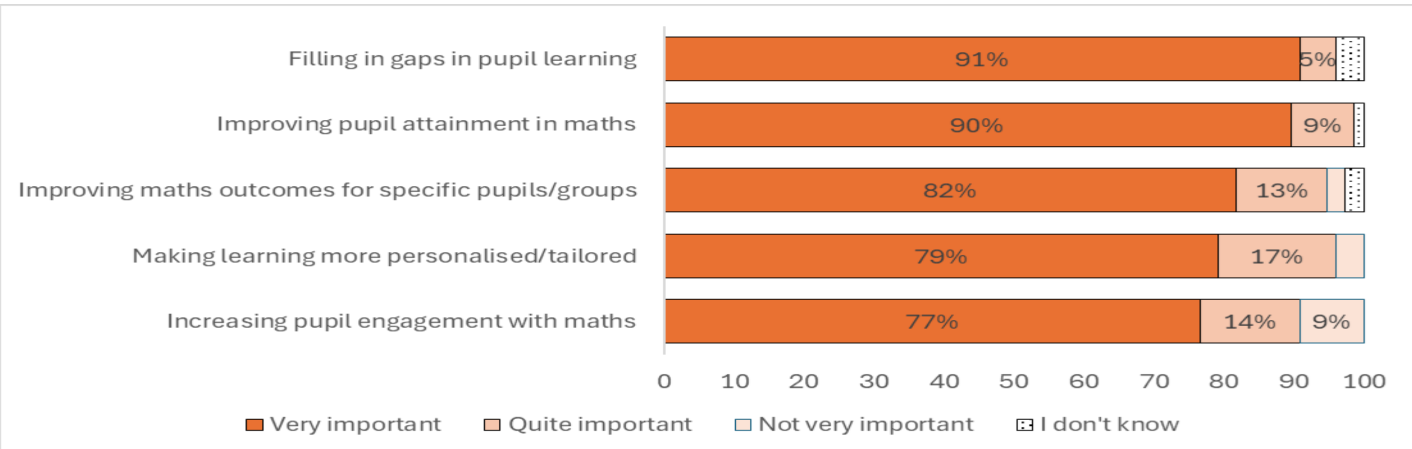
Figure 5: Teacher views on the problems which Maths-Whizz aims to address



Teacher baseline survey, n=257.

Similarly, school leaders' motivations for joining the trial align strongly with the intended approach and impact of Maths-Whizz, with over 90% of leaders agreeing that filling gaps in pupil learning and improving pupil attainment in maths were 'very important' or 'quite important' reasons for joining the trial (Figure 6).

Figure 6: Importance of different motivations for joining the Maths-Whizz trial



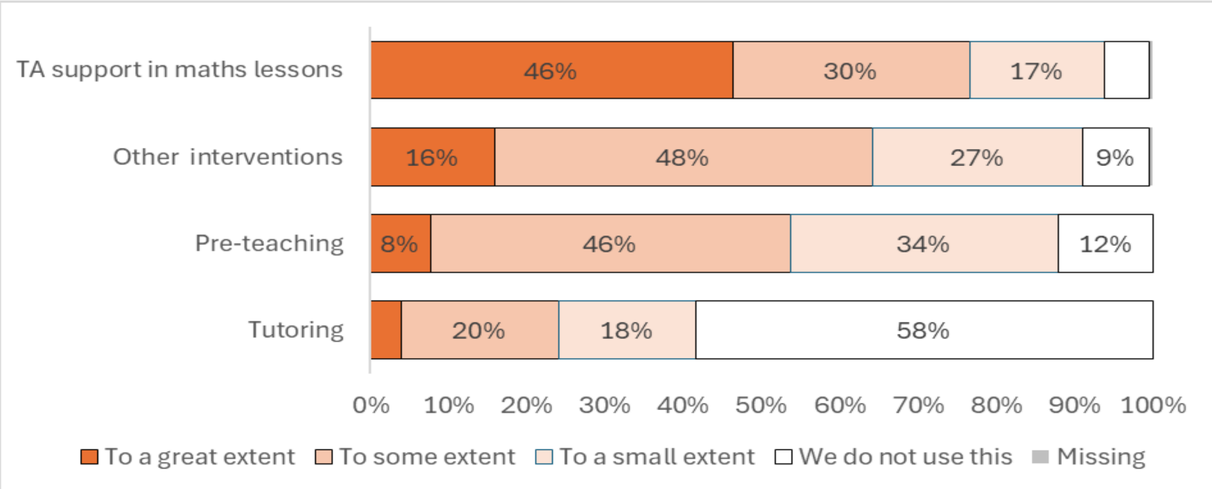
Leader endpoint survey, n=77.

Most teachers participating in the trial were highly experienced, with a mean teaching experience of 12 years at the start of the trial (baseline survey, n=257). Only a small proportion (15%) were early career teachers or trainee teachers, which is in line with the proportion of early career teachers in the primary workforce. At the start of the trial almost all teachers agreed they were confident in their knowledge of maths content (95%) and maths pedagogy (90%).

Usual practice (RQ5)

Before the trial, the most common forms of supplementary maths support were in-lesson support from teaching assistants (77% of teachers using to 'some extent' or 'a great extent'), structured intervention programmes (e.g. PiXL [Partners in Excellence] group sessions) (64%) and pre-teaching of maths content (54%). Relatively few teachers reported using tutoring (24%). Details are shown in Figure 7 below. In this context, 84% of teachers assigned to Maths-Whizz said the programme was 'very/moderately' different to their previous approaches to supplementary maths support.

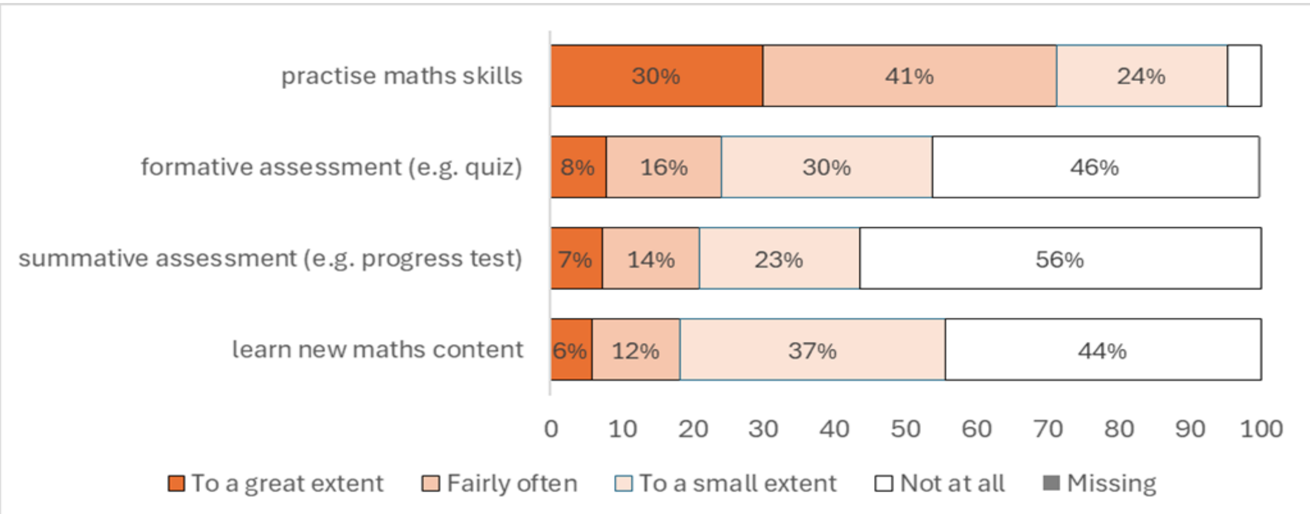
Figure 7: Usual practice in supplementary maths support



Teacher baseline survey, n=257.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

Considering use of digital/online tools for maths learning, most teachers (71%) commonly used these tools for practising maths skills, a feature which Maths-Whizz offers through its practice mode, in which pupils review and practice prior learning. The most popular tool reported by teachers for practising maths skills was Times Tables Rock Stars. However, less than a quarter of teachers commonly used online/digital tools for the primary purposes of Maths-Whizz: learning new maths skills (18%), which in Maths-Whizz is provided through the 'tutor mode'; formative assessment (24%); and summative assessment (21%). Details are shown in Figure 8 below.

Figure 8: Usual practice in using digital/online tools for maths learning



Teacher baseline survey, n=257.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

These findings suggest that Maths-Whizz was similar to usual practice in terms of providing practice and consolidation but showed strong differentiation from usual practice in using a digital/online tool to learn new maths skills and complete assessments. In addition, only a small majority of teachers (59%) agreed they were confident integrating technology into maths learning (27% neutral, 16% disagreed), suggesting that experience using Maths-Whizz could potentially support teachers to develop confidence in this area.

Comparing only the matched sample of teachers who completed both the baseline and endpoint surveys, there were some shifts[2] in usual practice in maths support during the trial which affected control (n=62–64[3]) and intervention (n=78) classes similarly. At endpoint, fewer teachers used pre-teaching to 'some extent' or 'a great extent' (decrease of 19 percentage points for control classes, and 12 percentage points for intervention classes). Fewer teachers used numeracy/maths interventions to 'some extent' or 'a great extent' (decrease of 13 percentage points for control classes, and 14 percentage points for intervention classes).

One shift in usual practice affected intervention classes but not control classes, namely the use of teaching assistant support in maths lessons. While control classes continued to deploy their teaching assistants to the same extent in maths interventions, the use of teaching assistant support declined (from 76% to 56%) in intervention classes (declined from 76% to 56% for intervention classes and remained at 75% for control classes). There is no data from the case study schools which would help to explain this change.

Leaders in case study schools confirmed they had not made significant changes to maths provision specifically for control classes, for example, because they wanted to be able to compare intervention and control class results within their school.

Together, these findings suggest that in general, classes in the control group did not compensate by introducing different forms of maths support in comparison with classes in the intervention group, and so provides a 'fair' comparator group for the Maths-Whizz intervention.

Fidelity (RQ1)

Core aspects of fidelity, as discussed in the logic model, include school leader oversight, use of Maths-Whizz tutor for 45–60 minutes per week during term time, the expectation that pupils will make three 'progressions' each week, and teachers' roles in supporting Maths-Whizz.

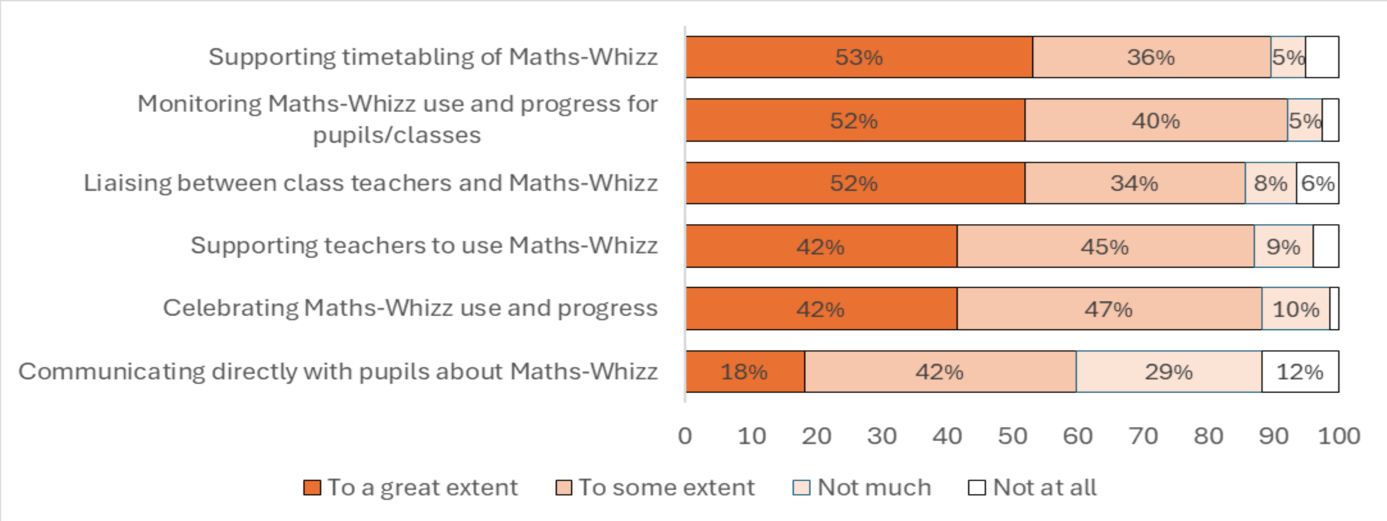
School leader oversight

Leader oversight of Maths-Whizz is a critical output in the logic model, which states that leaders are expected to 'strategically manage implementation and engagement with Maths-Whizz'.

The leader survey asked Maths-Whizz leads (typically the school maths lead) and headteachers about their role in supporting Maths-Whizz, including in training. Almost all leaders who responded to the survey (n=77) had attended the key leader meetings: the introductory webinar (n=75), implementation planning session (n=69), and termly reviews (n=68) (Figure 9). From survey open-text and interview responses, where leaders had not attended these meetings, it was commonly due to a change in role, for example, where a maths lead joined the school after onboarding and took up the role of Maths-Whizz lead.

[2] Shifts are reported for differences >10 percentage points between baseline and endpoint.
[3] The range refers to the number of teachers who responded to each item within this question.

Figure 9: Extent of leader involvement in different forms of Maths-Whizz support



Leader endpoint survey, n=77.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

The endpoint survey data shows strong engagement of leaders with the strategic oversight of Maths-Whizz in their school. Almost all leaders reported involvement 'to a great extent' or 'to some extent' in the key strategic tasks of supporting timetabling (89%) and monitoring Maths-Whizz use and progress (92%). In supplement to this strategic oversight, at least four-fifths of leaders were involved in day-to-day management of the implementation of Maths-Whizz in the school, for example, liaising between class teachers and Maths-Whizz, supporting teachers to use Maths-Whizz, and celebrating Maths-Whizz use and progress. A small majority of leaders (60%) also directly communicated with pupils about Maths-Whizz.

An artefact of the interleaved trial design, which deviates from the usual implementation of Maths-Whizz, is that only two of the year groups per school were using Maths-Whizz. This meant that in some cases, the designated Maths-Whizz lead (most commonly the school maths lead) was not using Maths-Whizz in their own teaching. In case study schools, Maths-Whizz leads who were not teaching Maths-Whizz classes reported that, although they had attended relevant training and reviews, not using it with their own classes made them less familiar with the platform, and less equipped to advise other teachers. This should be taken into consideration in the design of any future trials of Maths-Whizz, and by schools considering using Maths-Whizz with selective year groups.

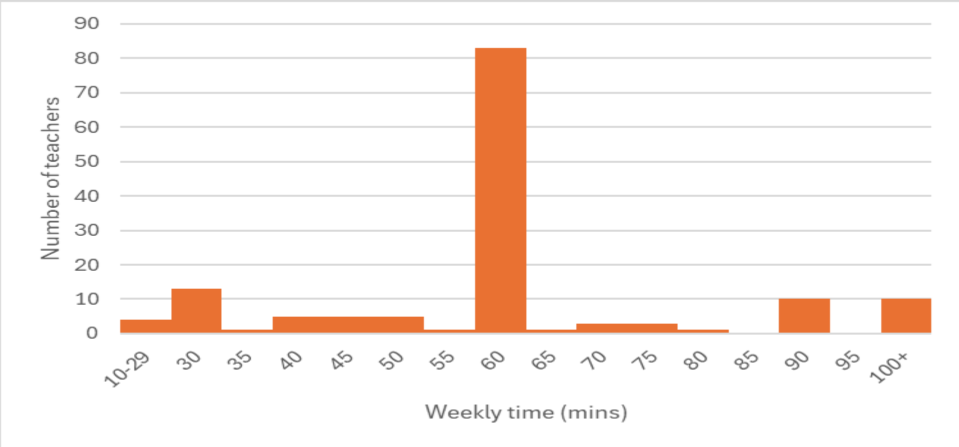
Supporting 45–60 minutes of tutoring per week

The key pupil input in the logic model is the use of Maths-Whizz tutoring for 45–60 minutes per week. This timetabling expectation was communicated in recruitment documents, in the onboarding training which all Maths-Whizz class teachers were expected to attend, and in the implementation plan documents which Whizz ESPs agreed with leaders in each school. In both observed sessions of onboarding training, Whizz Education staff strongly emphasised the importance of timetabling 60 minutes of Maths-Whizz tutor use within school time (training observations, August 2024 – September 2024).

The endpoint survey asked teachers the amount of class teaching time they allocated to Maths-Whizz during a typical week in the Spring Term and Summer Term. This period was chosen to reflect the allocation after allowing the Autumn Term for Maths-Whizz timetabling to be established.

Most teachers who responded to the survey were able to timetable Maths-Whizz in line with programme expectations (Figure 10). About three-quarters of teachers (73%) and leaders (79%) reported it was 'very easy' or 'moderately easy' to incorporate Maths-Whizz into the timetable. In line with this, over three-quarters of teachers (77%, n=111) reported allocating at least 60 minutes of Maths-Whizz teaching in a typical week, while 82% (n=122) of teachers reported allocating at least 45 minutes. A small proportion (7%, n=10) reported allocating 100 minutes or more, which is substantially higher than recommended, and might pose challenges for the rest of the teaching curriculum.

Figure 10: Minutes of class time allocated to Maths-Whizz in a typical week in the Spring Term and Summer Term



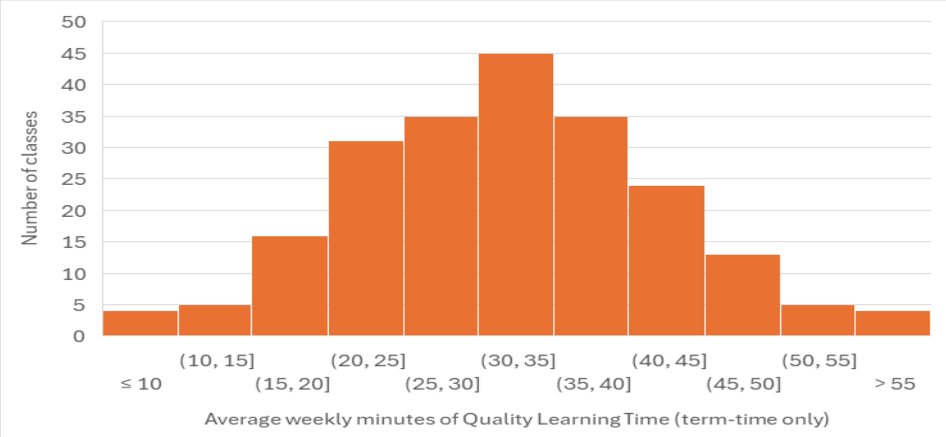
Teacher endpoint survey, classes which allocated time to Maths-Whizz (n=145).

Timetabling of Maths-Whizz within class time was a challenge for about a fifth of teachers. Almost a fifth of teachers (18%, 26 teachers across 14 schools) reported timetabling less than 45 minutes of Maths-Whizz teaching per week within class time. About half of these (n=12) used homework as part of their 'core' 45–60 minutes of Maths-Whizz, which might explain why most of these teachers still reported positive perceptions of Maths-Whizz's impact. Similarly, a fifth of teachers (20%) reported that timetabling challenges limited pupils' engagement with Maths-Whizz 'a lot', and about a fifth of teachers (19%) and leaders (21%) reported that it was 'not very easy' or 'not at all easy' to incorporate Maths-Whizz into the timetable.

Another data source for understanding the amount of tutoring time is the usage time which is logged by the Maths-Whizz platform for each pupil. This comprises the total time spent in Maths-Whizz focused on learning, across the 'tutor' and 'practice' modes. We expect the average pupil Quality Learning Time to be lower than the typical allocated class time (reported above) for several reasons. First, there will be some examples of zero time, for example if individual pupils are absent for a Maths-Whizz session, or a whole-class session is cancelled due to a trip or end-of-term activities. Second, allocated class time needs to allow for set-up and shut down, for example distributing and collecting devices and logging in. Third, the logged Quality Learning Time excludes any platform time spent on games and other activities which are not focused on learning. Figure 11 below shows the class-level distributions of Quality Learning Time. This is calculated as a weekly average for term-time only, in order to compare with the allocated class time.[4]

Figure 11: Class-level distribution of pupil Quality Learning Time logged in Maths-Whizz, based on term time

[4] Based on the dates between the initial Maths-Whizz assessment and endpoint testing for each pupil and subtracting eight weeks for school holidays. As the platform data does not distinguish between home use and in-class use, this will include any home use for classes who used Maths-Whizz at home, in term time and/or during school holidays. This means that the platform time is a slight overestimate of in-class use.



Maths-Whizz platform data, n=215.

Please note that the compliance analysis for the impact evaluation uses tutor time only to define compliant classes, while this IPE analysis includes both the 'tutor' and 'practice' modes of Quality Learning Time in Maths-Whizz. Based on the platform data, about a third of classes (35%) achieved the intended average Quality Learning Time of at least 45 minutes per week. This included 11% of classes with average term-time use above the expected 45–60 minutes, ranging from 61–91 minutes. Almost four-fifths of classes (78%) met a lower threshold of at least 30 minutes average Quality Learning Time per week during term time.

While the mean amount of class time allocated to Maths-Whizz was 63 minutes, the mean class-level Quality Learning Time (time that pupils spent engaged in 'tutor' or 'practice' modes in Maths-Whizz) was 40 minutes, assuming term-time use only. This suggests that, on average, only about two-thirds of timetabled time is translated into learning-focused time within the Maths-Whizz tutor. Given the centrality of Maths-Whizz Quality Learning Time as a mechanism for impact, increasing the translation of Maths-Whizz allocated class time to logged Quality Learning Time would be expected to support enhanced impact. Two potential mechanisms which would reduce logged tutor time and may need to be compensated for when allocating class time, are 'missed' weeks (e.g. due to school trips, or end-of-term timetable changes) and reduced efficiency of class time allocated to Maths-Whizz.

One factor, which is likely to influence the efficiency of Maths-Whizz time is how the class teaching time is scheduled across the week. While Whizz Education recommend using one or two sessions in a week, because this maximises tutor time and minimises set-up and shut down, schools are free to choose their own schedule and number of sessions and typically refine their approach over time (developer interview, December 2024). The most common models reported for the Spring Term and Summer Term were one (43%) or two (28%) sessions per week. However, 14% of teachers split Maths-Whizz time over four or more sessions, which may have reduced efficiency of the sessions.

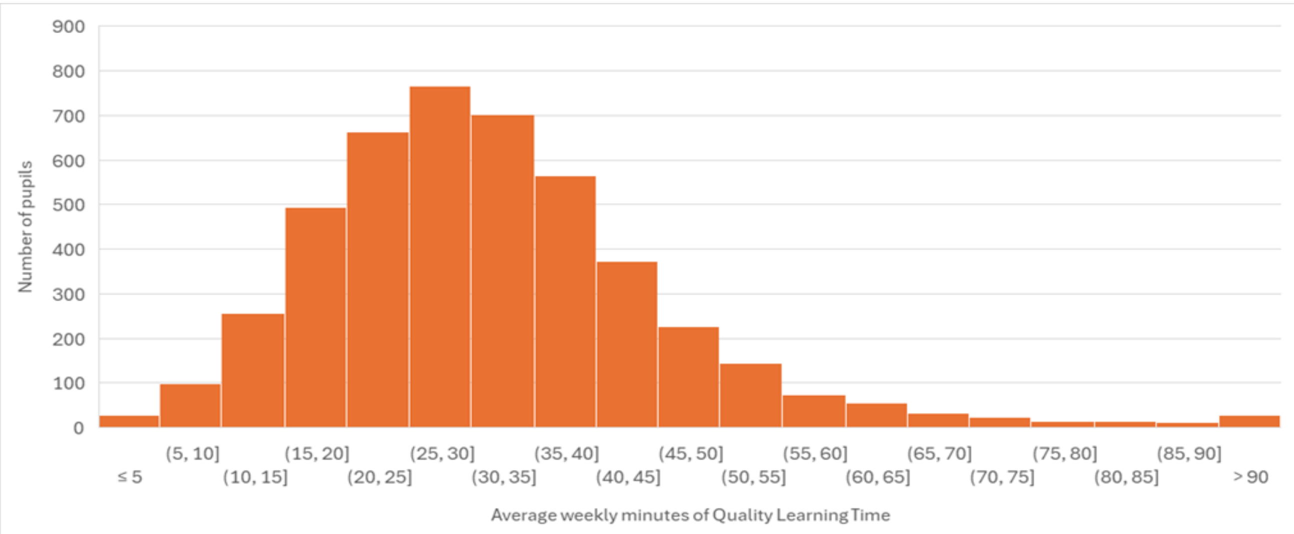
For example, teachers from one case study school described the logistical challenges of timetabling Maths-Whizz in the form of three shorter sessions across the week. Despite the planned timetabling of 60 minutes a week across three sessions, the classes involved did not reach the threshold of 30 minutes of Maths-Whizz usage per term week.

> *Teacher A: Trying to minimise that whole, like you say, unplugging....all [the devices], bringing them up, logging them in, getting set-up, starting, logging off, shutting down, putting them away, plugging them back in again.*
>
> *Teacher B: the children who probably need it the most are the ones that are not getting through it in the 20-minute time. And if they start on their quiz section and the 20 minutes is then up and we don't have the leniency to give them longer, it disappears and they lose it. And then they have to restart all over again.*
>
> *Teacher A: ...Some of them get very, very frustrated.*
>
> (Two class teachers from the same school, assigned to use Maths-Whizz)

As reported in the impact evaluation dosage analysis, when considering the whole period of Maths-Whizz use (including school holidays), the average weekly Quality Learning Time for pupils was 32 minutes. The distribution is shown in Figure 12 below.

Quality Learning Time was very similar for both FSM-eligible pupils (31.9 minutes) and other pupils (32.3 minutes). This suggests that FSM-eligible pupils were able to access the same amount of learning time as their peers. In all case study schools, staff were conscious of equity of access for their pupils. Some minimised differences in access by keeping Maths-Whizz within class time only and not sharing logins for home use. Others offered additional school-based time for pupils who were not able to access Maths-Whizz at home, for example, providing dedicated time during other lessons, or offering additional sessions before/after school.

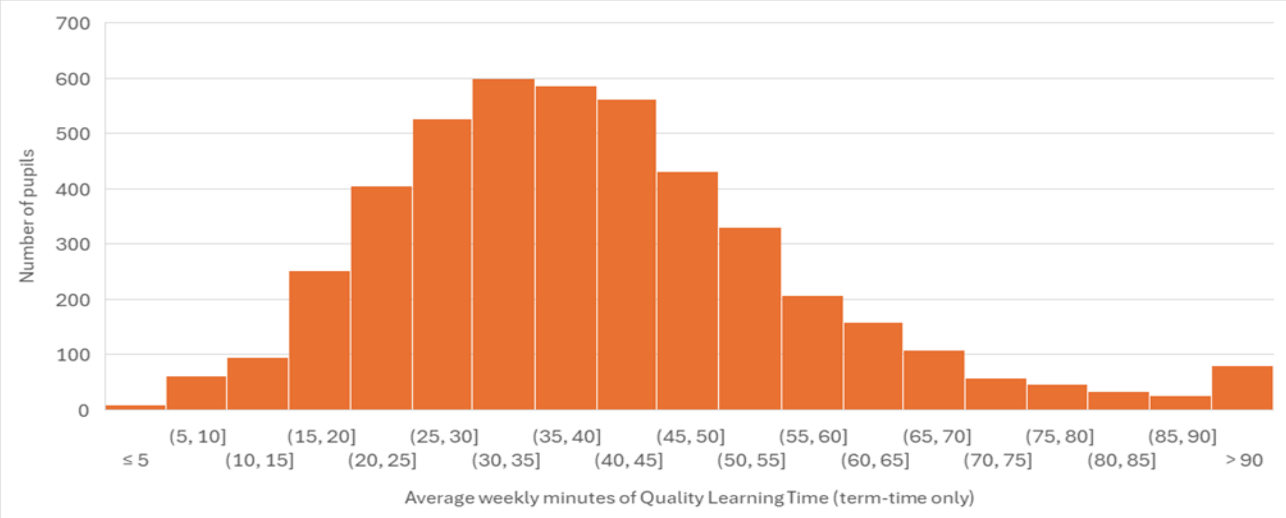
Figure 12: Pupil-level distribution of average weekly Quality Learning Time logged in Maths-Whizz (including holidays)



Maths-Whizz platform data, n=4,567.

Considering term-time use, Figure 13 below shows the pupil-level average Quality Learning Time adjusted for term-time only (subtracting eight weeks for holidays). The average term-time weekly Quality Learning Time was 40 minutes. About a third (32%) of pupils met the expected threshold of at least 45 minutes during term time, while 71% met the lower threshold of at least 30 minutes during term time.

Figure 13: Pupil-level distribution of average weekly Quality Learning Time logged in Maths-Whizz (term-time only)



Whizz platform data, n=4,567.

Reaching three progressions per week

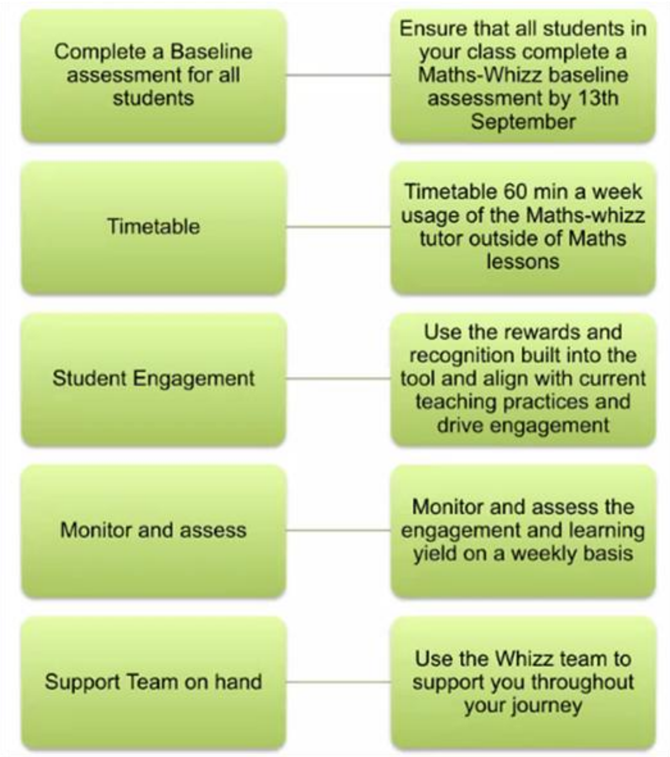
Pupils are set the goal of gaining three 'blue gems' each week by completing three new lessons in 'tutor mode'. Each lesson comprises a short tutorial, supported practice questions, and an assessment. In addition, pupils who gain three 'blue gems' from progressions in 'tutor mode', as well as three 'red gems' from practice and revision of prior learning in 'practice' mode, are admitted into the platform's weekly 'Hall of Fame' as a celebration and reward. In case study schools, both teachers and pupils were able to explain the expectations of gaining 'progressions' or 'gems', and teachers used this as a focus for their monitoring. Teachers displayed the Maths-Whizz visual progression track on the board during Maths-Whizz sessions so that pupils and the class teacher could see, monitor, and celebrate progress during the lesson.

From interviews, both school staff and the developer were aware that the red gems given for completing exercises on their prior learning had initially acted as a disincentive for a small minority of pupils to engage with tutor mode, which focuses on new learning. Unsurprisingly, pupils found their prior learning exercises easier and quicker to work through, compared with the new topics, and so prioritised collecting the red gems rather than the blue gems. In discussion with Whizz, teachers addressed this by asking pupils to complete their blue gems first (and monitoring this via the live progress reporting). This minor issue was resolved before the case study visits (March 2025).

Teachers' roles in supporting Maths-Whizz

The onboarding training sessions emphasised three ongoing roles for teachers: timetabling (see 'Fidelity' section above), monitoring pupil use and progress of Maths-Whizz, and building pupil engagement, for example, through in-platform and school-based rewards and recognition (Figure 14).

Figure 14: Project expectations shared with school staff during onboarding training



Training observations, August 2024 – September 2024

The logic model anticipated that, during the one-year trial period, teachers would use Maths-Whizz reporting data to ensure that pupils were using Maths-Whizz and completing progressions. Case study teachers described personal routines for this, for example, using the live screen progress map during Maths-Whizz class time to monitor and intervene in real time, checking reports via their phone, or using stickers to remind parents to encourage specific pupils. Based on the teacher endpoint survey, almost all teachers (93%) engaged with this process to some extent, with almost half of teachers (46%) doing this 'a lot'. Similarly, almost all teachers (88%) were using Math-Whizz data to reward pupils, with half of teachers (50%) doing this 'a lot'. From case study visits, rewards included the school systems (e.g. reward points) and tailored certificates printed from the platform, as well as small rewards (e.g. pencils, wristbands) and Maths-Whizz medals, which were provided to schools by Whizz Education.

Variation in implementation approaches (RQ1, RQ2, RQ6)

Whizz Education works with each school to develop a 'bespoke' implementation plan, which is tailored for the school context, while supporting each school to implement with fidelity through reaching the intended amount of tutor use and leader/teacher support. Variations in how schools choose to use Maths-Whizz are actively encouraged to ensure Maths-Whizz fits the school's goals and context. Key variations in implementation approaches seen across and within trial schools included the additionality of Maths-Whizz time, whether Maths-Whizz was used at home, and how teachers use Maths-Whizz, including integration with class teaching, and any challenges or barriers they experienced.

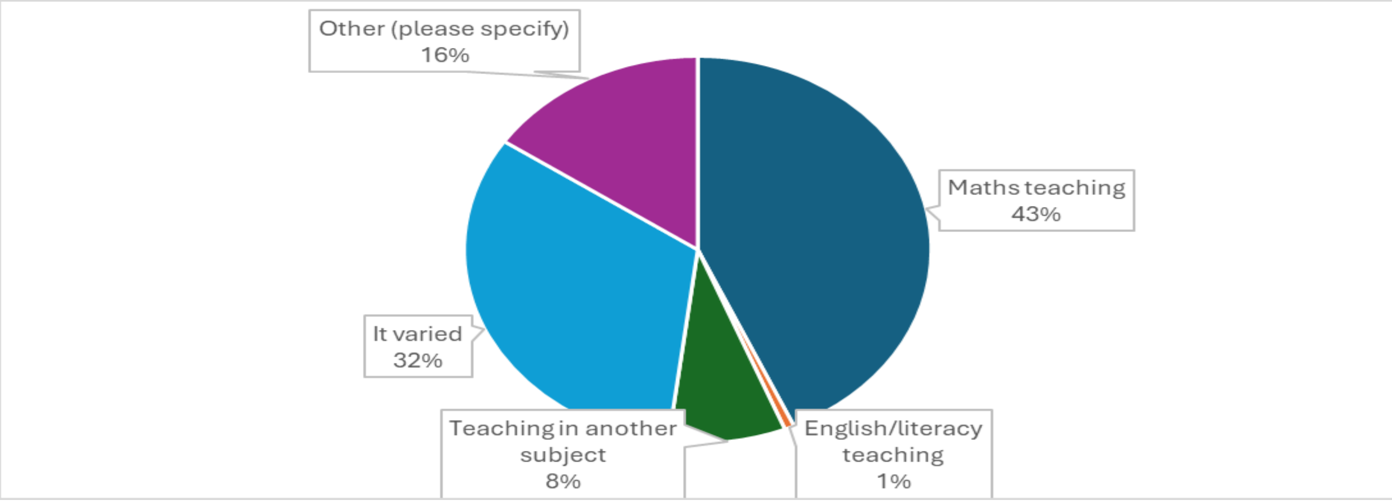
While implementation planning was done at school level, the detailed approaches to implementation varied within schools, across different classes and/or year group. For example, based on endpoint survey data from Maths-Whizz classes in 58 schools, in about half of these schools, classes took different approaches to what activity Maths-Whizz replaced (e.g. maths time or a subject) (52%, n=30). In over a third of these schools, classes had different approaches to home use (38%, n=22). Similarly, both the class time allocated to Maths-Whizz and the achieved usage time in Maths-Whizz varied substantively for different classes.

The variation within school cases meant that it was not feasible to conduct the planned qualitative comparative analysis, which systematically maps implementation characteristics to outcomes. Instead, teacher endpoint survey data was used to explore how implementation characteristics related to key perceived outcomes.

The additionality of Maths-Whizz

The expectation of the programme is that Maths-Whizz should be used in addition to timetabled maths lessons, rather than being a replacement for maths lessons.

Figure 15: The main activity which Maths-Whizz most commonly replaced in Spring Term and Summer Term



Teacher endpoint survey, n=148.

From the endpoint teacher survey, the most common main activity which Maths-Whizz replaced in the class timetable was maths teaching. This was reported by 43% of teachers, and most commonly by Year 5 teachers (54%) (Figure 14). In almost half of schools, Maths-Whizz time was replacing maths teaching time, rather than being additional curriculum time invested in maths. The total time spent learning maths by pupils in the intervention year groups in these schools may not be very different from the time spent on maths by control pupils, and this might have attenuated the impact seen in this trial. In the developer interview (December 2024), Whizz Education clarified that while they were aware of schools using time previously allocated to maths teaching for Maths-Whizz, for example, in the 'fifth lesson' of maths per week, the full maths curriculum was still being covered through teaching, and so Maths-Whizz was still considered supplementary to maths teaching. The same perspective was reflected by staff in case study schools who had repurposed some of their maths teaching time to Maths-Whizz. As these schools were implementing Maths-Whizz within the range of approaches which were acceptable to the programme developer, they are considered to be implementing with fidelity.

Home use of Maths-Whizz

Approaches to home use of Maths-Whizz varied substantially, including within schools. The most common approach was to leave home use at pupils' discretion (35%), while about a quarter of teachers reported that pupils were asked to use Maths-Whizz at home (28%) or conversely were asked not to use Maths-Whizz at home (26%). While Whizz Education onboarding training asked schools to timetable the 'core' time for Maths-Whizz within school time, a small minority of teachers (12%, almost all teaching Key Stage 2 year groups) asked pupils to use Maths-Whizz at home as part of their 'core' time. Among case study schools who were restricting use of Maths-Whizz to school, the two drivers for this were concerns about potential inflated progress from home use 'assisted' by parents or siblings, and concern about equity of technological access (e.g. devices, stable internet) at home. However, teachers at schools with an above-average proportion of FSM-eligible pupils (>24%) were slightly more likely than other teachers to encourage home use (44% compared with 36%). In case study schools encouraging home use, teachers were aware of some pupils that did not have home access, and provided additional school time for these pupils, either before/after school or within the school day. Across the trial cohort, usage time was similar for FSM-eligible pupils and their non-FSM-eligible peers (see below), suggesting that schools managed to ensure access for FSM-eligible pupils.

The developer reported that for a very small number of the pupils who used Maths-Whizz at home, weekly usage became very high (developer interview, December 2024). Whizz Education had a system to manage these cases, to ensure that Maths-Whizz use remained at a sustainable level and to manage screen time for pupils. The pupil was tagged within the platform, and Whizz Education, in consultation with the school, set a 'usage cap' to ensure more sustainable and balanced use of Maths-Whizz. This issue occurred for specific pupils, rather than being spread across a whole class or school.

Implementation characteristics and perceived outcomes

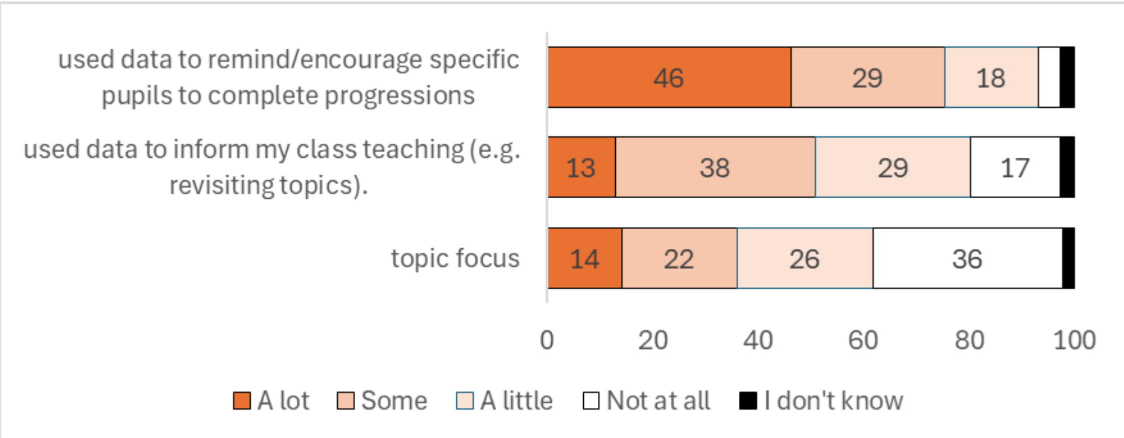
Based on endpoint survey data, teachers who reported any use of Maths-Whizz at home were more likely than those reporting no home use to report a perceived strong positive impact on achievement in school maths assessment (29% of teachers who required home use, 29% of teachers who let pupils decide, and 18% of teachers who asked pupils not to use Maths-Whizz at home). There were no substantive differences in perceiving a strong positive impact of Maths-Whizz on achievement in school maths assessment by whether Maths-Whizz replaced some maths teaching time. These findings should be treated with caution as not all teachers responded to the endpoint survey.

Based on teacher and leader interviews, decisions about timetabling, what Maths-Whizz displaced and home use were made thoughtfully, considering their school context and how Maths-Whizz would work best for their pupils. Some schools had refined their approach during the year based on teacher and pupil experience and/or Maths-Whizz reporting data. This suggests that the flexibility of implementation encouraged by Maths-Whizz is an important element of the programme.

Integrating Maths-Whizz and maths teaching

Many teachers were beginning to integrate Maths-Whizz more directly into their maths teaching, by directing Maths-Whizz teaching through the 'topic focus' feature or using Maths-Whizz data to inform class teaching (Figure 16). Similarly, most case study teachers were aware of 'topic focus' as a feature of Maths-Whizz. From the logic model, these features were expected only in second and subsequent years of use. This suggests that trial teachers were developing sophisticated use of Maths-Whizz even within the one-year trial.

Figure 16: Teachers' use of Maths-Whizz features to integrate tutoring and class teaching



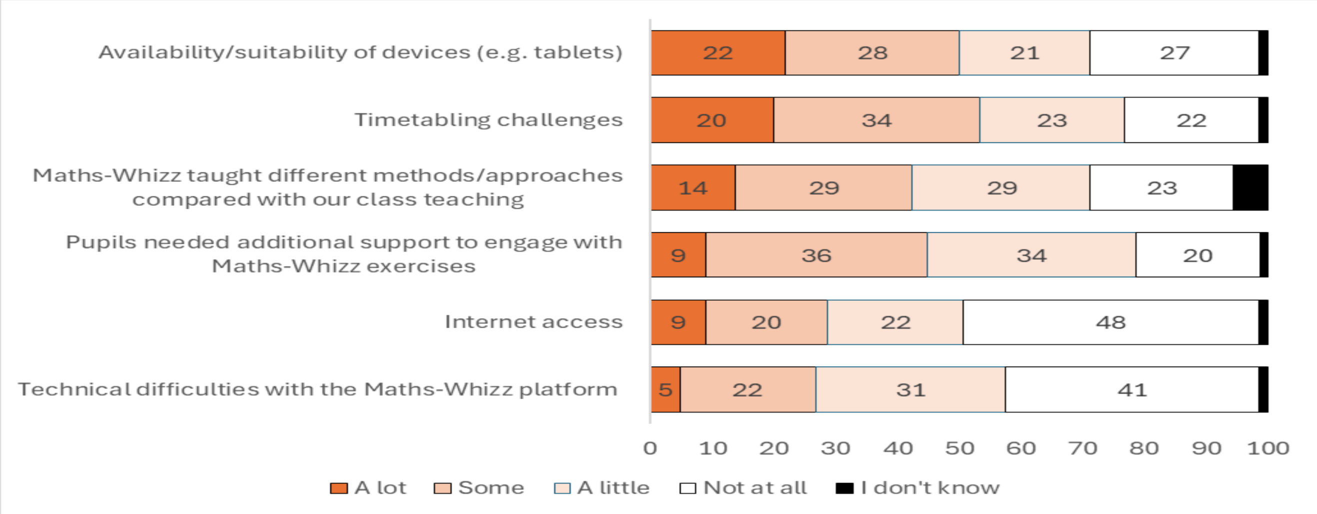
Teacher endpoint survey, n=148.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

In addition to the platform, Maths-Whizz teachers were also given access to the Maths-Whizz teacher resource, which provides curriculum resources to support class teaching and homework. Most teachers (80%) reported using this, with 13% of teachers using it 'a lot'. While teachers of control classes also had potential access to the Maths-Whizz teacher resource, almost all of these teachers (92%) reported that they did not use it, with most of the remaining teachers using it only 'to a small extent'. Therefore, we do not consider this to be a significant source of contamination in the trial.

Barriers and challenges affecting implementation

Barriers and challenges to engaging with Maths-Whizz, which were commonly mentioned in the case study visits were added to the teacher endpoint survey to assess their prevalence. These are shown in Figure 17 and described below.

Figure 17: Extent to which specific challenges limited class engagement with Maths-Whizz



Teacher endpoint survey, n=148.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

The most commonly reported technological challenge was device availability and suitability (50% of teachers reported this limited engagement 'a lot' or 'some'). About a quarter of teachers were also limited 'a lot' or 'some' by issues with internet access (29%) or technical difficulties with the Maths-Whizz platform (27%), with a large overlap between these groups, though the most common response from teachers was that they were not limited by these issues 'at all' (48% and 41% respectively). Based on case study schools, a key issue with device availability and suitability was agreeing timetabling for use when devices were shared by multiple years or classes, and ensuring devices were appropriately charged.

Pedagogically, just under half of teachers reported that engagement was limited 'some' or 'a lot' by some pupils needing additional support (45%) beyond the scaffolding and feedback provided within Maths-Whizz. Interviewed teachers who experienced this explained that it was challenging to support all the individual pupils who needed help with either the technology or the maths learning. In their usual maths teaching, teachers could intervene by providing scaffolding and support for a small group or the whole class together, but in Maths-Whizz pupils were working on individually tailored activities, so teacher support needed to be individual.

Just under half of teachers reported that engagement was limited 'some' or 'a lot' by Maths-Whizz teaching different methods and approaches compared with class teaching (43%). Interviewed teachers who had experienced this explained that it was challenging for pupils when Maths-Whizz taught them a different process for a calculation from the one they had learned, and that some pupils struggled with following an unfamiliar method and became 'frustrated'.

Responsiveness (RQ2)

Experience of Maths-Whizz

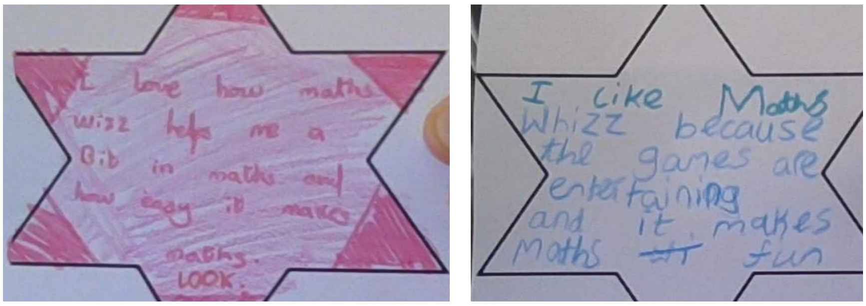
Almost all staff reported that that they were satisfied with the Maths-Whizz programme (rated 'very/moderately' satisfied by 95% of leaders and 93% of teachers in the endpoint surveys) and that pupils enjoyed using Maths-Whizz (rated 'very/moderate' by 100% of leaders and 96% of teachers). In case study schools, many teachers reflected that pupils 'looked forward' to their Maths-Whizz sessions. Where teachers offered pupils the option of additional Maths-Whizz time, for example, after completing classwork, or in 'free choice' time, pupils responded positively and chose Maths-Whizz above other options. A minority of case study teachers reported that some pupils did not enjoy using Maths-Whizz. The two reported reasons for this were negative experiences with Maths-Whizz (e.g. due to device, internet, or platform challenges, or use of short sessions where pupils could not make progressions), and, for some pupils, prior dislike of maths.

In focus groups, case study school pupils completed a 'stars and wishes' activity to identify anything they liked about Maths-Whizz, and anything they disliked or would like to change (Figures 18 and 19). Across all case study schools, most pupils in focus groups were positive about Maths-Whizz. Some pupils made broad positive comments about Maths-Whizz or identified their favourite topic or lesson, these topics varied individually. More specific comments related to the 'gaming' nature of Maths-Whizz, and the provision of adaptive learning and support.

Many pupils enjoyed the gamified nature of learning in Maths-Whizz, for example, the game style, sound effects, and timers, acquiring 'gems' as they progressed, and the 'Hall of Fame' available to pupils who achieved the weekly target of gems. Reflecting this, pupils across case study schools described Maths-Whizz as 'games you play' that helped with maths, rather than as a maths learning platform. 'Fun' was commonly used to describe Maths-Whizz. Some pupils commented positively on the additional 'play zone' within the Maths-Whizz platform, including choosing pets and other items to buy from their earned 'coins', the animations, and pupils being able to play games when they 'need a break'.

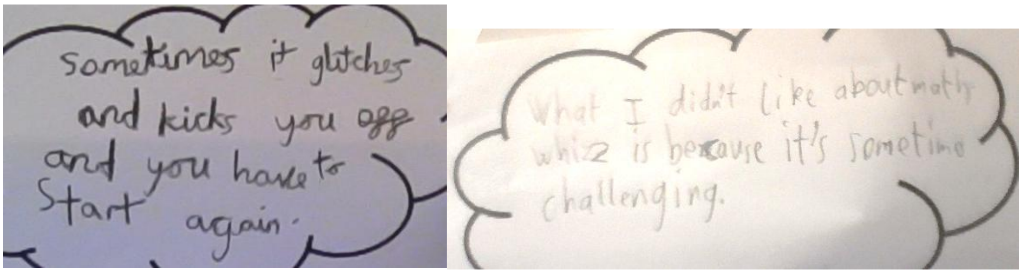
Most pupils were aware of the adaptive learning approach used by Maths-Whizz, for example, sharing that Maths-Whizz gives you practice in the lessons where you need more support. They were positive about the scaffolded learning, appreciating that Maths-Whizz would provide videos and other help when you did not know something, and 'takes things slow'. Similarly, pupils in one class appreciated being able to move directly to lessons pitched for them: *'if it is easy, you can jump ahead'* (pupil who used Maths-Whizz).

Figure 18: Examples of positive pupil comments from focus groups



Where pupils shared negative comments or suggested improvements, some were general or focused on maths topics they individually disliked. Some pupils did not enjoy the level of challenge and difficulty, which is a direct consequence of the adaptive learning, which asks pupils to work at the edge of their current capability. While no pupils mentioned disliking red gems, some disliked the blue gems. This may reflect teachers' comments that the pupils noticed that blue gems, which require new learning, were harder and took longer than the red gems. Similarly, one pupil commented that *'I always get the lessons I hate'* (pupil who used Maths-Whizz), which is likely a consequence of the adaptive tutor assigning topics the pupil has not yet mastered. Pupils in some schools disliked that they had lost progress due to technical issues, though this was not frequent for the pupils who experienced it. Other requests were for more features in the 'play zone', for example, a wider range of pets or games.

Figure 19: Examples of pupil negative comments from focus groups



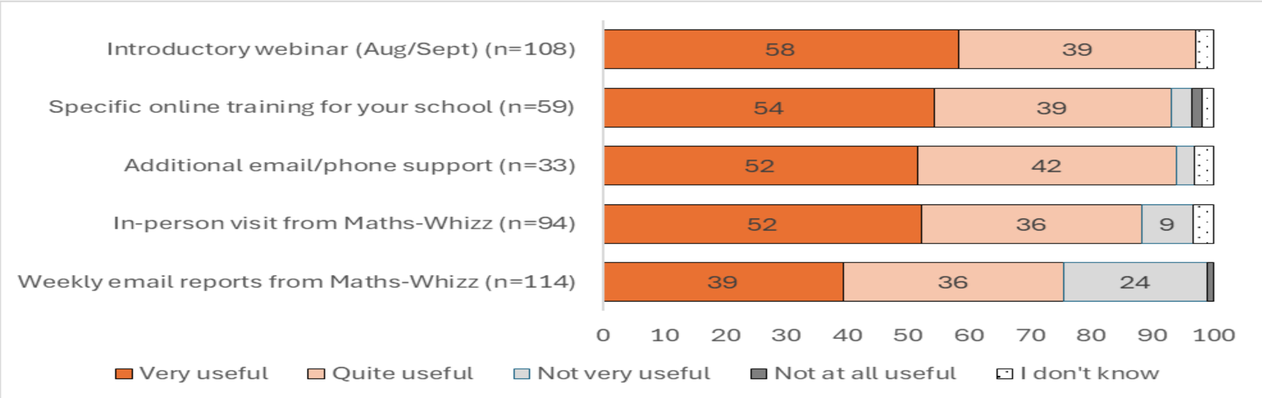
Whizz Education support for teachers and leaders

Whizz Education's enhanced support model includes substantial training and support for implementation. This includes an onboarding webinar, a leader planning session to agree a contextualised implementation plan for how the school will use Maths-Whizz, school-specific online training, an in-person visit, termly reviews, as well as ad hoc support by phone and email. Another key element of support is a weekly email sent to teachers and leaders, which provides an overview of class and pupil progress. All leaders (100%) and almost all teachers (90%) were 'very' or 'moderately' satisfied with the support they received from Maths-Whizz.

In the endpoint survey, school staff were first asked which training and support activities they had engaged with and were then asked to rate the usefulness for each activity they had engaged with.

The number of teachers who reported engaging with each activity varied, from 33 who reported asking for additional support by email or phone, to 114 who were aware of receiving weekly email reports. Almost all teachers (88–97%) found the specific forms of training and support they reported engaging with 'very useful' or 'quite useful' (Figure 20). Most teachers (75%) found the weekly email reports from Whizz Education 'very useful' or 'quite useful', while almost all the remaining teachers found them 'not very useful'. In case study schools where Maths-Whizz time was allocated to a single weekly session (which was the most common timetable model, reported by 43% of surveyed teachers), teachers reported primarily tracking progress live on screen during class time, rather than via emails, which may explain the lower rating.

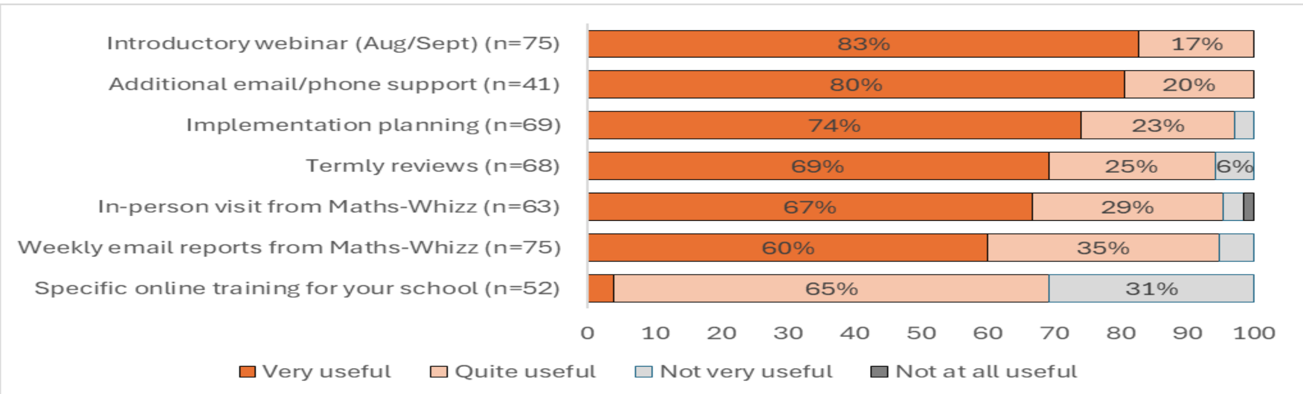
Figure 20: Teachers' views on the usefulness of training and support from Whizz Education



Teacher endpoint survey, n=33–114.[5]
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

Similarly, almost all leaders (88-97%) found most forms of training and support they engaged with 'very useful' or 'quite useful' (Figure 21). Again, the number of leaders who reported engaging with each activity varied, from 41 who reported asking for additional support by email or phone, to 75 who remembered attending an introductory webinar and were aware of receiving weekly email reports. The exception was school-specific online training, which leaders most commonly rated 'quite useful', while teachers most commonly rated it as 'very useful'. This may reflect that some Maths-Whizz leads were teaching control groups, so not using Maths-Whizz directly with their own classes.

Figure 21: Leaders' views on the usefulness of training and support from Whizz Education



Leader endpoint survey, n=41–75.[5]
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

[5] Based on the number of teachers/leaders responding to the endpoint survey who reported accessing this support.

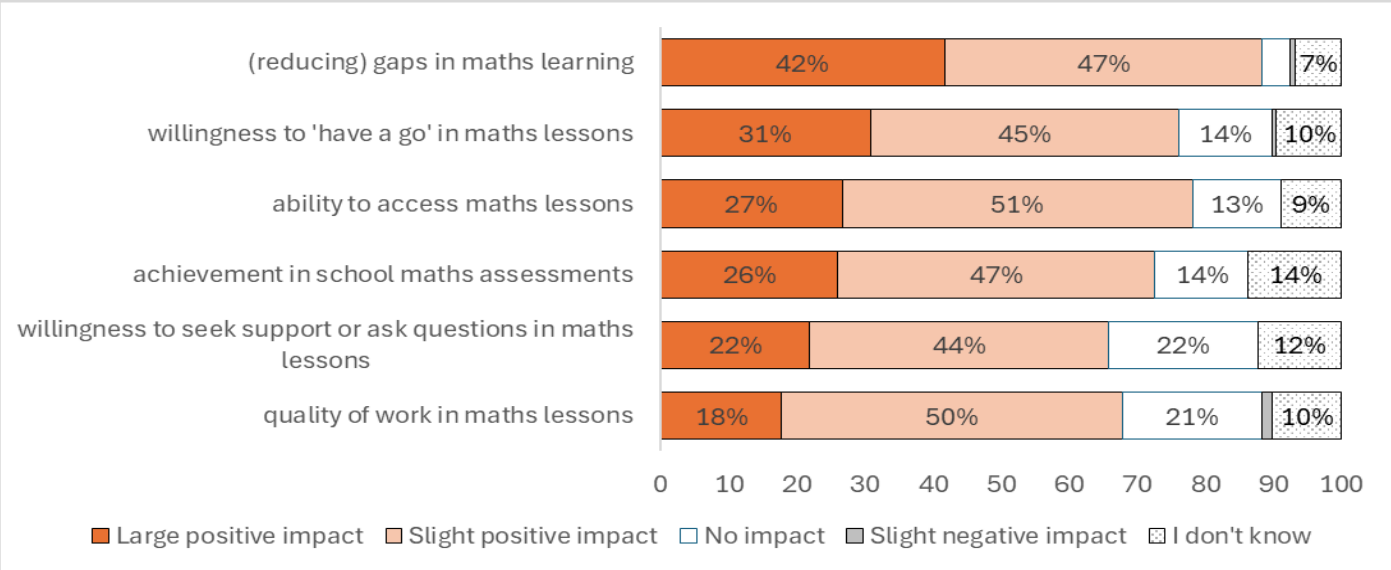
Staff in all case study schools, including those with lower usage or neutral/negative views of the Maths-Whizz programme, were very positive about the support they received from Whizz Education. In particular, staff praised the quick responses and helpfulness of the Whizz Education team.

Perceived impact, mediators, and moderators (RQ3, RQ4)

Perceived impact for pupils

Teachers' perceived impact of Maths-Whizz on pupils is shown in Figure 22. More than three-quarters of teachers reported positive impacts on reducing gaps in maths learning (88%), pupils' ability to access maths lessons (78%), and their willingness to 'have a go' in maths lessons (76%). Most teachers reported positive impacts on pupils' achievement in school maths assessments (73%), the quality of work in maths lessons (68%), and pupils' willingness to seek support or ask questions in maths lessons (66%). Less than a quarter of teachers (4–22%) reported no impact on these outcomes and almost no teachers (0–1% per outcome) reported negative impacts.

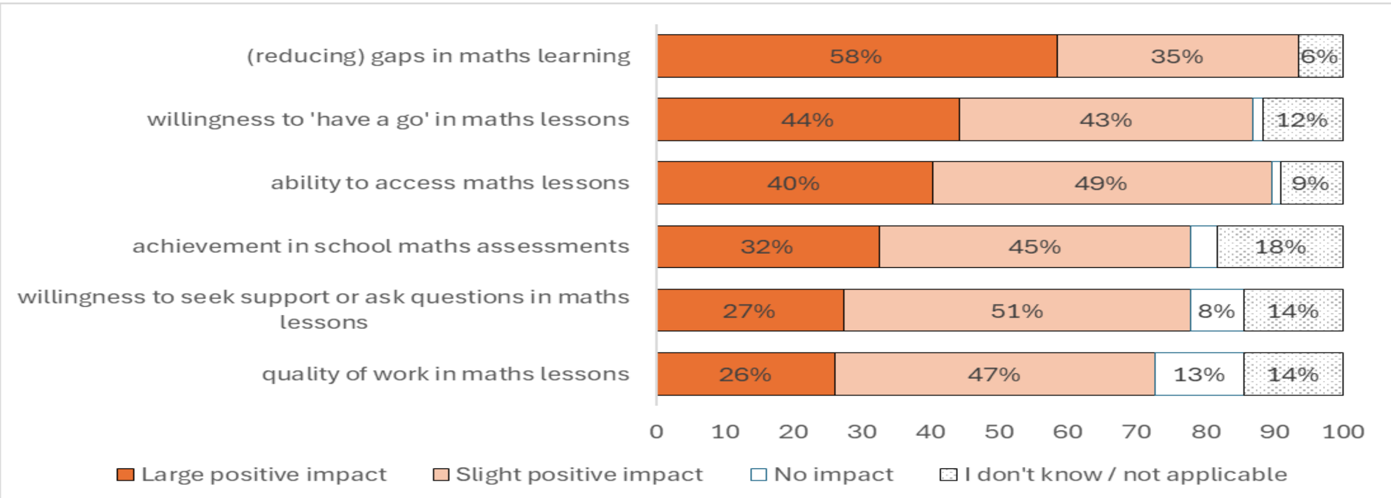
Figure 22: Teacher views on perceived impact on pupils



Teacher endpoint survey, n=148.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

Leaders' perceptions of impact (Figure 23) were even more positive than teachers' perceptions. The order of outcomes from most to least positive was the same, suggesting a consistent perspective across school staff that Maths-Whizz is most effective at reducing gaps in maths learning.

Figure 23: Leader views on perceived impact on pupils



Leader endpoint survey, n=75.
Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

In focus groups, pupils were asked about their 'biggest change' from using Maths-Whizz. Pupil responses were inductively grouped into three themes:

- **Learning outcomes.** Pupils learned more maths, or improved in specific topics (e.g. multiplication, reflection).
- **Learning process.** Pupils increased their participation and learning behaviours in maths, such as asking and answering questions in class, or feeling confident to try an activity.
- **Change in attitude towards maths.** Pupils now 'liked' or 'loved' maths, or looked forward to maths lessons (beyond those where they used Maths-Whizz). This is consistent with our findings that Maths-Whizz had a small positive effect (though not significant) on pupils' enjoyment of maths.

This variety reflects the intended pupil outcomes in the logic model, and reflects the perceived impact reported by teachers. Many pupils linked the changes they reported to the personalised support provided by Maths-Whizz:

> *I was not that good at maths but now I am, because it gives you practice when you aren't good at the lesson* (Pupil writing in focus group)

Pathways for improved pupil outcomes

The IPE design anticipated three key pathways by which Maths-Whizz might contribute to improved outcomes in maths (RQ4). The first hypothesised pathway was that pupil learning with Maths-Whizz could directly improve maths outcomes. This was the most prevalent pathway mentioned by interviewed staff from case study schools. Among case study schools who achieved at least 30 minutes of average usage, staff saw pupil progress within the Maths-Whizz platform and felt that using Maths-Whizz was filling gaps in pupil learning. Similarly, 42% of surveyed teachers reported a strong positive impact on filling gaps in pupil learning. At the time of case studies (after about six months of Maths-Whizz use) there were mixed views on whether this had translated into a broader impact on maths learning, for example, a positive impact on school maths assessments. While some teachers were seeing Maths-Whizz progress mirrored in their lessons and assessments, others had not seen evidence of this transfer:

> *They could do it all on Maths-Whizz, and are they pulling it into lessons, maybe sometimes, but then when maybe you're given a question that's worded differently...do they know how to answer it, are they able to understand that? Probably not.* (Maths-Whizz lead)

The second hypothesised pathway was that using Maths-Whizz could improve pupils' engagement with maths teaching, which would then enhance the impact of class teaching. Some surveyed teachers perceived a strong positive impact on engagement with maths teaching (31% reported a strong positive impact on pupils being willing to 'have a go' in maths lessons, 27% on pupils' access to maths teaching, and 22% on pupils' willingness to seek support or ask questions). Reflecting this, teachers/leaders in case study schools had mixed views on whether the high levels of pupil engagement they saw within Maths-Whizz time was transferring to their maths lessons.

The third hypothesised pathway was that changes in class teaching (e.g. based on Maths-Whizz data) improved outcomes in maths. Most case study teachers said they had not substantially changed their class teaching based on Maths-Whizz data, although there were occasional examples. For example, one teacher had noticed from the Maths-Whizz data that many pupils in her class were working on a specific topic in Maths-Whizz, and encountering similar challenges, and adapted class teaching to address that. In line with the case study findings, a small minority of teachers (14%) said that Maths-Whizz had a strong positive impact on how they approached maths teaching with their class.

When case study teachers/leaders were asked about how Maths-Whizz supported improved learning for pupils, the most common suggestions were that Maths-Whizz provides tailored consolidation and practice to support pupil fluency, and that in cycling through topics it acts as spaced retrieval practice.

> *It's just really complemented what we've been doing. And in some areas, consolidating the learning. So for example, when we've done shapes...for some children who didn't quite grasp 3D, 2D...they were doing it*

on Maths-Whizz as well. I think it helped them to understand certain differences between 3D, 2D as well as knowing those different shapes…Whereas if it wasn't there, would they have had that opportunity in that moment? Probably not, because we would have come to the end of that topic and we've moved on to something else like multiplication and division. (Teacher)

[Children] were having problems just retrieving that information, using it in their working memory. I think they're the children that we've seen the most success with, that they're now able to practice things a little bit more and then be able to think, oh yeah, and bring that forward then into their maths lessons…I guess it's the more you do something is it? Whereas if you've not looked at your times tables for months, that knowledge sort of falls away. (Maths-Whizz lead)

Differences in pupil outcomes

Teachers were asked about the differential benefit of Maths-Whizz for specific pupil groups, including FSM-eligible pupils, pupils with Special Educational Needs and Disabilities (SEND), different attainment levels, and gender (Table 28).

Table 28: Perceived differential impact for specific pupil groups

Group	Benefited more than other pupils	Benefited equally with other pupils	Benefited less than other pupils
FSM-eligible pupils	11%	57%	5%
Pupils with SEND	26%	36%	24%
Lower-attaining pupils	34%	41%	14%
Higher-attaining pupils	34%	48%	5%
Female	5%	63%	1%

Teacher endpoint survey, n=148. The other response categories 'Mixed/I don't know' and 'My class doesn't have any pupils in this group' are not reported here, therefore, percentages do not sum to 100%.

For each pupil group asked about, the most common response was that pupils in this group benefited equally with other pupils (varying from 36–63% depending on the group). This suggests that teachers found Maths-Whizz to be broadly beneficial, including for FSM-eligible pupils. In case studies, no staff spontaneously mentioned FSM-eligible pupils as benefiting differentially from Maths-Whizz, but when asked specifically about FSM-eligible pupils, they reported that the programme worked well for them. The most varied response was for pupils with SEND, with about a quarter of teachers each perceiving that SEND pupils benefited more (26%) or less (24%) than other pupils.

Leaders were also asked about differential benefits for specific pupil groups. For each pupil group, fewer than 10% of leaders thought this pupil group benefited less than other pupils, reinforcing that school staff found Maths-Whizz to be broadly beneficial. Leaders were more likely than teachers to report a differential positive benefit for FSM-eligible pupils (29% of leaders, 11% of teachers), pupils with SEND (41% of leaders, 26% of teachers), and lower-attaining pupils (45% of leaders, 34% of teachers).

Case study interviews shed light on the mixed survey responses for SEND, which reflect the heterogenous nature of SEND and the accommodations and provision available in specific schools. Where Maths-Whizz was perceived to be working well for pupils with SEND, staff highlighted pupils being able to work independently and at their own pace, not being singled out, the use of headphones to block out classroom noise, the programme's ability to read out loud, and reduced barriers relating to social when using Maths-Whizz. Where Maths-Whizz was perceived to be working less well for pupils with SEND, staff highlighted that some pupils could not access the reading required, which was an issue for schools which did not have access to headphones,[6] or that pupils needed individualised staff support to engage with Maths-Whizz, which was challenging to provide in a whole-class context.

[6] Headphones were recommended for using Maths-Whizz but were not a requirement for joining the trial.

At least a third of teachers perceived that Maths-Whizz had more benefit for higher-attaining and lower-attaining pupils (34% each). From staff interviews, case study schools had signed up for Maths-Whizz with the expectation that it would support their lower-attaining pupils and agreed that it had done so. Some were surprised that Maths-Whizz had supported other groups of children so well, for example, higher-attaining pupils working at greater depth. Staff framed this as an unexpected benefit of Maths-Whizz. One possible interpretation for a differential benefit for both higher- and lower-attaining learners is that the individualised nature of Maths-Whizz was particularly beneficial for pupils who were further from the typical attainment and pace of progress of their class.

Perceived impact for staff

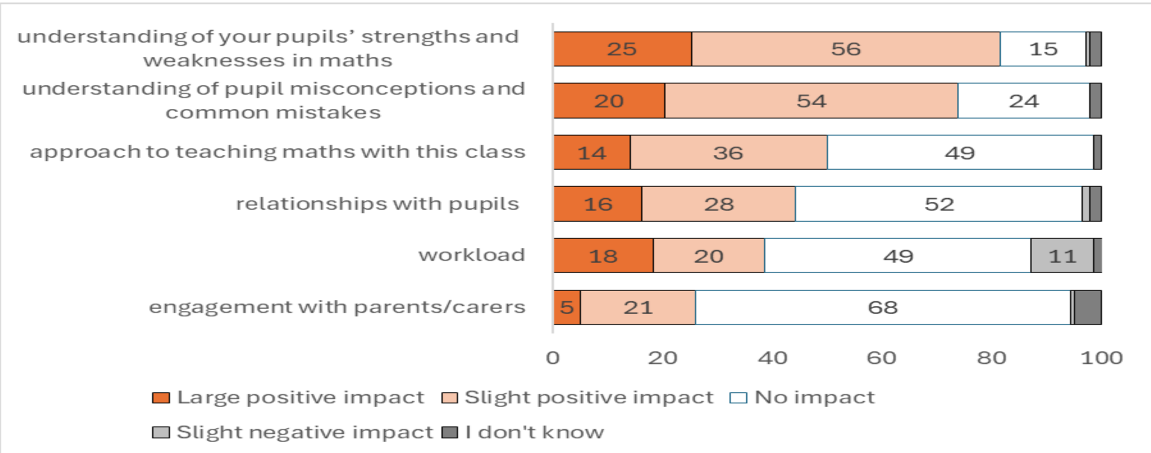
Interviewed staff from case study schools were asked about any impacts of Maths-Whizz for staff and the school. Impacts identified in case study data were then included in the teacher endpoint survey, alongside impact on workload, to understand their prevalence.

Within maths teaching, one interviewed teacher described that engaging with the Maths-Whizz reporting data, specifically pupil and class profiles, had given them a better understanding of their pupils' strengths and weaknesses. Similarly, for another teacher, watching pupils interact with Maths-Whizz, including the diagnostic feedback offered by the platform, had developed their understanding of common misconceptions and mistakes within the topics they were teaching. Another teacher identified that Maths-Whizz had influenced their approach to class teaching.

Teachers also identified positive relational impacts from using Maths-Whizz. Some teachers felt that Maths-Whizz had encouraged more positive and celebratory relationships with their pupils. The clear visual progress, platform messaging and certificate functions, and Whizz's provision of a medal and small rewards, supported them to encourage, cheerlead, and reward pupils as they worked with Maths-Whizz. Similarly, across most case study schools, staff reported that pupils would approach them in the school grounds to tell them about their progress and appreciated direct platform messages from the lead. Similarly, one teacher reported improved engagement with parents/carers, from championing Maths-Whizz and sending reminders via school messaging systems and in-person at school pick-up and drop-off.

Based on the survey results (Figure 24), the most prevalent of these outcomes for Maths-Whizz teachers were increased understanding of pupils' strengths and weaknesses (81% large or slight positive impact), and of common misconceptions and mistakes (74%). Importantly, most teachers reported that using Maths-Whizz had a positive impact (39%) or no impact (49%) on their workload, although about one in ten teachers (11%) reported a slight negative impact on their workload. From surveyed teachers, the mean additional time for preparing and monitoring Maths-Whizz was 13 minutes per week, with 24% of teachers reporting 0 minutes of additional time, and only 5% of teachers reporting more than 30 minutes of additional time per week.

Figure 24: Perceived impact for teachers



Teacher endpoint survey, n=149.

Note: Where categories are combined in the text, they may not reflect the sum of categories due to rounding.

Cost

Costs are based on full market costs for the delivery model used in the trial for three years, i.e. access to Maths-Whizz for two of the year groups within the school. Whizz Education is working on approaches to lower access costs further. Whizz Education is piloting whether automation-driven delivery enables Maths-Whizz access to be provided at a lower price. In addition, at the time of reporting, Whizz Education is currently offering a contribution towards student access for UK state schools, resulting in pricing of £12 per student annually and reducing to £10 per student if a multi-year commitment is agreed.

Maths-Whizz has a fixed cost per pupil which provides unlimited access, i.e. 60 minutes per week has the same cost as 30 minutes per week. This is a potential benefit compared with human tutoring, where costs scale with tutoring time.

School-level subscription costs: We have calculated the school-level subscription costs based on the UK state school pricing for the 2025/2026 academic year onwards. This represents the most up-to-date costing and incorporates a recent reduction in pricing. We have assumed purchase of the 'enhanced' support model in the first year, as tested within the trial, followed by purchase of the 'essentials' support model in the second and third years, for ongoing light-touch support.

Number of pupils: The mean number of pupils per trial school (n=83) was calculated based on the number of pupils allocated to Maths-Whizz at randomisation (n=5,233) divided by the number of schools (n=63), rounded to the nearest integer. For this calculation, paired infant and junior schools are treated as a single school.

Additional costs: The majority of schools did not report additional costs. We have not included any additional costs in the primary cost model but have run a sensitivity check including median additional costs for schools who did report these.

Table 29: Cost of delivering Maths-Whizz

Item	Year one	Year two	Year three	Cost	Total cost over three years	Total cost per pupil per year over three years
Annual Maths-Whizz subscription (school base cost – year one 'enhanced' model)	✓			£1,875	£1,875	£7.53
Annual Maths-Whizz subscription (school base cost – year two and year three 'essentials' model)		✓	✓	£495	£990	£3.98
Annual Maths-Whizz subscription (per pupil cost)	✓	✓	✓	£25 per pupil	£75 per pupil	£25.00
Total					£2,865 + £75 per pupil	£36.51

Table 30: Cumulative costs of project A (assuming delivery over three years)

	Year one	Year two	Year three
Maths-Whizz Intelligent Tutoring	£47.59	£30.96	£30.96

Sensitivity checks

We ran sensitivity checks to test the robustness of this cost estimate. The resulting costs per pupil per year (calculated over three years) were in the range of £26 to £49.

The main driver for differences in costs is the number of classes/year groups using Maths-Whizz. For a whole-school subscription to Maths-Whizz (Reception to Year 6), assuming average class sizes,[7] **the cost per pupil per year ranges from £26 for a four-form entry school (761 pupils) to £30 for a one-form entry school (190 pupils).**

Staff from 29 schools reported additional costs via the trial lead/leader survey. A total of 26 schools reported costs for additional technology (range £50 – £4,000, median value £263), and nine schools reported costs for additional staffing (range £250-£10,000, median value £1,000). Six schools reported both types of cost. No schools reported other types of additional cost. The estimated cost per pupil per year for a school with £263 of start-up technology costs is £37.57, while the estimated cost per pupil per year for a school with £1,000 per year of recurring staff costs is £48.56.

Core costs for trial schools

Schools participating in the trial paid a reduced total subscription of £500, for access for the 2024/2025 academic year, due to subsidy from Whizz Education. This equates to a cost per pupil per year of £12.65.

Additional resources

Using Maths-Whizz requires a digital device (tablet or computer) per pupil and good web connectivity. This was an eligibility criterion for the trial, checked as part of the onboarding process, and so does not represent an additional cost in the context of the trial. Looking at the wider population of schools in England, from Year 5 TIMSS data (Richardson *et al.*, 2025), 43% of pupils had teachers who reported that digital devices were available during maths lessons, and that use of digital devices was 'not at all' limited by 'not enough access'. This is an indicative estimate of the proportion of schools in England who would be able to use Maths-Whizz without incurring additional device costs. Key limitations of this estimate are that it included teachers who reported that pupils shared devices, did not consider technical specifications, and uses data from summer 2023.

The minimum technical specifications **are:**

- Chrome 70+/MS Edge 90+/Safari 12+/Firefox 78+/Samsung Internet 11+.
- Minimum 7.9" screen (recommended for usability reasons).
- WebGL-enabled graphics card and browser.
- 2GB RAM.
- 0.5 Mbps internet connection per concurrent pupil.

The additional time commitments for schools, (i.e. beyond class teaching time), are detailed below, based on data from Whizz Education. As for the cost analysis, this assumes enhanced service for the first year of the programme (as tested in the trial) and the 'essentials' model for the second and third years of the programme.

All interviewed teachers confirmed that they supervised their classes while working with Maths-Whizz, so we assume that no contact time was saved for teachers. However, it is possible that teachers may reduce the time spent in lesson planning, preparation, and marking for Maths-Whizz, in comparison with the lesson it replaced. We did not collect time data on this, so have not included it in our estimates below. However, as reported in the IPE results, importantly, most teachers reported that using Maths-Whizz had a positive impact (39%) or no impact (49%) on their workload.

[7] In 2024/2025, the average class size was 26.2 in Reception/Key Stage 1, and 27.9 in Key Stage 2.

Table 31: Start-up and recurring costs for Maths-Whizz

Element	Description	Cost type	Who	Time per year (hours)	Total time over three years (hours)
Training	Onboarding webinar in first year	Start-up	All teachers	1	1
Training	School-specific training and visit in first year	Start-up	All teachers	2	2
Implementation planning and review	Implementation planning meeting at start, and three termly reviews across the first year	Start-up	Leader(s)	2	2
Teacher planning / support time	Weekly support time (in addition to class time)	Recurring	All teachers	8.5[a]	25.5

[a] *Based on the mean weekly time for planning and support reported in the teacher endpoint survey (13 minutes), and assuming a 39-week school year.*

Sustainability

Despite the positive perceived impact of Maths-Whizz, less than half of trial schools (40%, n=25), have continued to subscribe to Maths-Whizz for the current academic year (2025/2026) at the current market rate. Leaders in case study schools, including those who wanted to continue with Maths-Whizz based on the perceived impact for their pupils, indicated budget as the key constraint which would influence their decision to re-subscribe. This suggests that to ensure continued impact, sustainability is a key challenge to address before scaling.

Conclusion

Table 32: Key conclusions

Key conclusions
1. Pupils allocated to receive Maths-Whizz made one additional month's progress in maths, on average, compared to pupils who were not allocated to receive the programme. This finding has high security.
2. Among children eligible for FSM, those allocated to receive Maths-Whizz made two additional months' progress in maths, on average, compared to FSM-eligible pupils who were not allocated to receive the programme. These results may have lower security than the overall findings because of the smaller number of pupils.
3. Pupils allocated to receive Maths-Whizz reported a small positive effect on their mathematical self-efficacy and enjoyment of maths compared to pupils who were not allocated to receive the programme. This was supported by teachers' feedback that intervention pupils 'looked forward' to using Maths-Whizz. Pupils were also broadly positive about the programme and particularly enjoyed the gamified learning on Maths-Whizz.
4. On average, pupils engaged with the Maths-Whizz platform for around 32 minutes per week, with around 25 minutes spent in the tutor mode learning new content and eight minutes in the practice mode reviewing and consolidating content. Higher usage in both modes was associated with better maths attainment with time in the practice mode showing the strongest association.
5. Several features of Maths-Whizz might have contributed to this impact, particularly the larger impact for FSM-eligible pupils. These include adaptive teaching, dynamic feedback mechanisms (real time, adaptive responses that continuously adjust questions, difficulty, and support based on a student's performance in order to personalise learning), self-paced individualised learning experiences, and making gamified learning accessible, all of which make learning more accessible and equitable for disadvantaged pupils. The stronger impact for FSM-eligible pupils suggests that this platform may be particularly adept at identifying and addressing gaps in learning, which are more prevalent among FSM-eligible children.

Impact evaluation and IPE integration

Evidence to support the logic model

Maths-Whizz aims to address the twin problems of a high range of attainment and individualised learning gaps in primary maths classes, by providing individualised adaptive tutoring. Almost all trial teachers reported these problems in their previous years' maths classes, suggesting these are relevant issues to address. The logic model for Maths-Whizz describes the expected inputs, outputs, and outcomes for leaders, teachers, and pupils.

School leaders engaged in strategic oversight of Maths-Whizz, for example, by engaging with implementation planning for their school and termly reviews. Within the one-year trial, the expected outcome was that leaders were committed to and capable of implementing adaptive tutoring, which is demonstrated by the high proportion of compliant classes (91%, which contained 95% of pupils). The longer-term outcome for leaders, that they have a sustainable strategy and capability to continually improve maths learning outcomes, may be limited by the lower proportion of trial schools who continued to subscribe to Maths-Whizz for 2025/2026 (40%), which leaders reported was primarily due to financial constraints.

Most teachers were able to timetable sufficient class time for Maths-Whizz, although this was challenging and/or limited for about a fifth of teachers. Almost all teachers used reporting data to encourage pupils to use Maths-Whizz and make progressions. There was also evidence that some teachers were using more advanced features of Maths-Whizz, which are not considered 'core' in the first year, including directing Maths-Whizz to specific focal topics and/or using Maths-Whizz data to inform class teaching.

Pupils successfully completed initial assessments and sustained Maths-Whizz tutoring across the year. While 82% of teachers reported timetabling at least 45 minutes per week, with an average allocation of 63 minutes, pupils' average usage time within Maths-Whizz during term time was about two-thirds of this, at 40 minutes per week. As the dosage analysis found that learning time in both tutor and practice mode was associated with changes in maths attainment, the reduced average dosage for pupils is likely to have weakened impact on maths attainment. The efficiency of class time allocated to Maths-Whizz is a key line of enquiry for further research.

The IPE shows perceived short-term outcomes from using Maths-Whizz, including progression, enjoyment, and efficacy in pupils using Maths-Whizz, and filling individual learning gaps. The impact evaluation showed a positive effect on maths attainment and a small positive effect on self-perception and enjoyment in maths. The IPE shows a small perceived impact on pupils' engagement with maths lessons, including their ability to access lessons, willingness to seek support, and quality of work. The transfer of outcomes such as enjoyment and self-efficacy from Maths-Whizz use into broader maths learning is a key line of enquiry for further research.

The logic model hypothesised that Maths-Whizz may be particularly beneficial for FSM-eligible pupils. This was supported by the impact evaluation, which found a positive differential impact on the maths attainment of FSM-eligible pupils compared with their non-eligible peers. One hypothesised mechanism for this differential benefit was the higher prevalence of low attainment in FSM-eligible pupils. However, the impact evaluation findings did not support this hypothesis, as there was not a differential impact for low-attaining pupils compared with their peers. The IPE found that FSM-eligible pupils had similar Maths-Whizz usage time to their peers, suggesting that schools' strategies to achieve equitable access were working well, including for schools which allowed or encouraged home use. This is reflected in the view of most surveyed teachers, that FSM-eligible pupils benefited equally or more from Maths-Whizz, compared with other pupils.

Interpretation

The impact of the Maths-Whizz Intelligent Tutoring System on pupils' maths attainment (0.076; 95% CI: 0.02, 0.13) is consistent with existing evidence, which shows that Edtech interventions focused on maths have a small but positive effect on pupils' attainment (Cheung and Slavin, 2013; Lewin *et al.*, 2019). Our findings are also consistent with the small effect of intelligent tutoring systems (ITS) (effect sizes ranging from 0.01 to 0.09) on pupils' maths attainment (Steenbergen-Hu and Cooper, 2013). The evidence from research on intelligent tutoring systems and, more broadly, Edtech interventions also suggests that short, focused interventions (5–10 weeks or one term) used frequently (e.g. three times per week) are particularly effective at improving attainment, whereas sustained technology use over longer periods (e.g. one academic year as in this trial or longer) is less effective at improving attainment (Higgins, Xiao and Katsipataki, 2012; Steenbergen-Hu and Cooper, 2013). We note, however, that the finding about the duration of technology use, particularly intelligent tutoring systems use, is based on a meta-analysis where only three of the 96 effect sizes relate to primary education, which is of relevance to this study.

The range of effect sizes seen for technology-based interventions indicates that it is not technology use *per se*, but the pedagogical underpinning of technology use (i.e. how the technology is used to support teaching and learning) that drives impact (Higgins, Xiao and Katsipataki, 2012). In fact, the recommendations in the EEF guidance report on the use of technology to support pupil learning use effective pedagogy as the basis for technology use (Stringer, Lewin and Coleman, 2019). The recommendations focus on first understanding how the technology will improve teaching and learning and then using technology to improve: i) modelling and examples; ii) pupil practice; and iii) assessment and feedback. The implementation planning exercise prior to the start of Maths-Whizz delivery allowed each school to understand how they could use Maths-Whizz to improve teaching and learning in their school and develop a bespoke plan for implementing the programme in their school. The Maths-Whizz platform uses various effective pedagogical approaches including modelling, practice, assessment, and feedback. Pupils complete an initial assessment at the start of the programme to identify gaps in their understanding and prepare a customised learning journey. As pupils' complete lessons on the platform, they are first taught the topic of the lesson following which they complete some interactive problems on that topic. When pupils get stuck on a problem, the platform helps them by showing them how to approach the problem. The 'Topic Challenge' mode on the platform allows pupils to review, practice, and consolidate previously covered content. The platform also provided pupils with feedback when they completed practice tests at the end of each lesson and provided actionable insights to teachers on pupils' progress.

Findings from this trial suggest Maths-Whizz has a greater impact on the maths attainment of disadvantaged pupils (as measured by FSM eligibility) (0.14; 95% CI: 0.06, 0.22). This is an unusual finding that is inconsistent with existing evidence that Edtech (and intelligent tutoring systems) interventions have smaller effect sizes for disadvantaged pupils than larger school-based populations (Steenbergen-Hu and Cooper, 2013; Haßler *et al.*, 2025). The stronger impact of Maths-Whizz on FSM pupils suggests that this platform may be particularly effective at identifying and addressing gaps in learning which are likely to be more prevalent among FSM children because they have higher rates of absence (The Sutton Trust, 2024).

Practitioner feedback has highlighted specific features of Edtech intervention that are particularly relevant for making learning more accessible and equitable for disadvantaged pupils (Haßler *et al.*, 2025). Prominent examples included tools that use adaptive teaching and dynamic feedback mechanisms, promote self-paced or individualised learning experiences, or making gamified learning accessible in school, all of which are key features of Maths-Whizz. This finding that Maths-Whizz has a stronger effect for FSM pupils is particularly impactful when viewed in conjunction with our finding from the IPE that close to half the schools in the trial implemented Maths-Whizz during maths teaching time. Therefore, it appears that Maths-Whizz can improve maths outcomes for FSM pupils without significantly taking away time from other subjects for these pupils. Another important factor was the planning of implementation in a way that did not exacerbate existing disadvantages. Whizz Education recommends that 'core' Maths-Whizz time is timetabled in school time. While a small minority of teachers in the trial asked pupils to use Maths-Whizz at home as part of their 'core' time, this was not a common approach to the implementation of the programme. Although Maths-Whizz shows promise to reduce the attainment gap for disadvantaged pupils, it is important to note that not all disadvantaged pupils are the same. In fact, the distribution of attainment for disadvantaged pupils has a similar range to that of their peers (DfE, 2015). Research shows that the impact of disadvantage is higher for non-English as an Additional Language (EAL) pupils compared with EAL pupils (Treadaway, 2017). Similarly, the number of years a pupil has been eligible for FSM (i.e. duration of disadvantage) impacts attainment cumulatively (Treadaway, 2017). Future research can unpick some of these potential drivers of impact for disadvantaged pupils.

Maths-Whizz had a positive impact on pupils in each different year group, with the largest impact on Year 3 pupils (0.13; 95% CI: 0.03, 0.24) and the smallest impact on Year 5 pupils (0.02; 95% CI: -0.08, 0.12). Impacts for Year 4 (0.08; 95% CI: -0.01, 0.18) and Year 2 pupils (0.07; 95% CI: -0.11, 0.26) fell in between these two. However, the CIs for all year groups were overlapping suggesting that the apparent differences between year groups may be due to sampling variation and the programme has a broadly similar impact across all four of the year groups. The evidence for the variation of the impact of Edtech interventions by the age of learners is also inconclusive, with some studies finding variation with age and others not (Li and Ma, 2010; Cheung and Slavin, 2013). We did not find any evidence of difference in dosage of Maths-Whizz experienced by the different year groups, thereby ruling this out as the reason for the apparent differences in impact between year groups. Implementation of the intervention was good in all year groups. As the Maths-Whizz platform provides tutoring to Reception to Year 8 pupils, it would be useful for future research to explore how and why the programme might impact younger and older children's ability to learn. Based on our best estimates of impact, our recommendation is that schools with limited budgets focus their delivery of Maths-Whizz to Year 3 pupils, for whom the programme appeared to have the greatest impact.

Compliance at the class level was defined either in terms of time spent on the platform in the tutor mode (at least 30 minutes per week) or in terms of the number of tutor exercises attempted (at least three attempted per week). While 91% of all intervention classes were compliant on the basis of this definition, only 38% of these classes were compliant on the basis of time spent in the tutor mode on the platform. Most were compliant on the basis of the number of tutor exercises attempted. Overall, pupils in compliant classes made the equivalent of two additional months' progress in maths attainment, suggesting that Maths-Whizz can benefit all pupils in a class when it is implemented in this way.

Limitations and lessons learned

This was an efficacy trial for which the developers recruited schools and delivered the programme. We anticipated that the interleaved design where all schools in the trial would receive access to Maths-Whizz in two of the year groups would be attractive to schools during recruitment. As Whizz Education recruited the schools to the trial following a detailed conversation with senior leaders to ascertain eligibility and feasibility of and support for implementation, these schools were more likely to be open to the approach used by the developers, compared to the broader population of all schools in England. Additionally, in this trial, we have tested the enhanced support model offered by Whizz Education, which involves detailed implementation planning, teacher onboarding and training, formative and continuous assessment, periodic data and progress check-ins, an annual review and ongoing support. Whizz Education also offer an essentials support plan that includes continuous assessment and real-time reporting, online training materials, and live support whose efficacy has not been tested here. Therefore, the generalisability of the results to all schools should be made with some caution.

A potential source of contamination was from the same teacher teaching both control and intervention year groups. All intervention year group teachers were given access to the Maths-Whizz Teachers Resource, which provided them with lessons, worksheets, enrichment sheets, and classroom resources that they could use for maths teaching. Although teachers were instructed to not use these resources with control pupils, there is a chance that this occurred. In fact, 92% of teachers with control classes reported no use of these resources, while almost all others said they used these to 'a small extent'. We deem the risk of contamination to be quite low as even if the teachers used these resources with control pupils, they did not have access to the Maths-Whizz platform, which is the key input for the primary pathway to impact for pupils.

Maths-Whizz is designed to be used in addition to timetabled maths lessons rather than a replacement for them. However, the IPE revealed that in close to half the schools in the trial Maths-Whizz time replaced some maths teaching time. The developers clarified that they were aware that some schools were using maths teaching time to timetable Maths-Whizz sessions. However, they were of the view that as the full maths curriculum was being covered in maths teaching, the Maths-Whizz sessions were still additional. While this may, in fact, be accurate, the total time spent learning maths by pupils in the intervention year groups in these schools may not be very different from the time spent on maths by control pupils and this might have attenuated the impact seen in this trial. We suggest that the additionality of Maths-Whizz, including whether allocating more time to maths has consequences for learning in other subjects be explored in greater detail if this programme proceeds further along the EEF pipeline to effectiveness trial.

School-level attrition was low with only one out of the 66 schools (1.5%) withdrawing during the main stage of the trial and not completing the endpoint assessment. Pupil-level attrition was higher, with 13% of pupils not completing the endpoint assessment. A key contributor to pupil attrition in both intervention and control groups was the withdrawal of one school from the trial. Other reasons for missing pupil endpoint data were pupil absence on test day, pupils leaving the school, withdrawal of pupils from testing by the teacher because of additional needs that made accessing the Star Maths assessment untenable and internet connectivity issues, which caused some tests to be paused midway. There was slightly a difference in attrition for intervention and control pupils (14% vs 13%). It is unlikely that there is bias due to missing data because reasons for missing were unrelated to randomised group and our missing data analysis showed that the main effect size was robust to sensitivity checks.

From the perspective of analysis of 'interleaved' trials, there were three key learnings. The first is the treatment of year group in modelling. Whereas our approach to the primary analysis model in the Statistical Analysis Plan was to consider year group only as a random effect, during analysis, this resulted in an error (Schwendel, Welbourne and Sahasranaman, 2025). This was the result of pupils of different year groups being measured on the same scale, with younger pupils tending to score lower on average, leaving very little residual variation within year groups for the random effects to capture. Therefore, we readjusted the primary analysis model to consider both the fixed effect of year group to account for these differences between younger and older pupils, and the random effect, which could then isolate the year group differences within schools. The second key learning concerned how Hedges' g was calculated in a trial where outcomes were measured across multiple year groups. In practice, Hedges' g was derived using a pooled SD across eight groups (intervention and control within each year group), rather than across just two groups (overall intervention and control) as originally specified in the Statistical Analysis Plan (Schwendel, Welbourne and Sahasranaman, 2025). This was considered necessary because the sample comprised pupils from four different year groups. Combining scores from multiple year groups would artificially inflate the pooled SD relative to a study involving a single year group. This could create the counterintuitive situation where effect sizes for individual year groups appeared larger than the overall effect size when all year groups were combined without adjustment. Using the eight-group pooled SD avoids this issue by controlling for between-group variation, producing Hedges' g estimates that are more realistic and comparable with those calculated separately for each year group. As a robustness check, we also calculated Hedges' g using age-standardised scores. However, for consistency with other statistics presented in the report—which are based on Unified Scaled Scores—we have chosen to report Hedges' g throughout using Unified Scaled Scores and a pooled SD calculated across the eight groups. The third learning relates to partitioning of ICCs in models where both year group and class are included as random effects. In the proposal we had assumed that roughly one-third of the variance would remain accounted for at the school level, in fact the school-level variance was much smaller than this with the bulk of the variance being accounted for at year group level and class level. This suggests that the two-level models that are most frequently employed in the EEF evaluations may not adequately reflect the true structure of the data.

Future research and publications

Future research can examine the drivers of impact of the Maths-Whizz programme for disadvantaged pupils. As suggested previously, this can begin with an examination of the EAL status of disadvantaged pupils as this has been shown to moderate the impact of disadvantage on attainment. Understanding how pupils' age moderates the impact of Maths-Whizz on maths attainment is another potential line of enquiry for future research. A future multi-year trial spanning Years 3 and 4 or even Years 2, 3, and 4 would help us examine whether the effect of Maths-Whizz is additive. The transfer of outcomes such as enjoyment and self-efficacy from Maths-Whizz use into broader maths learning and whether these outcomes are sustained in the absence of access to Maths-Whizz are other key lines of enquiry for further research.

References

Adelson, J.L. and McCoach, D.B. (2011) 'Development and psychometric properties of the math and me survey: measuring third through sixth graders' attitudes toward mathematics', *Measurement and Evaluation in Counselling and Development*, 44(4), pp. 225–247. Available at: https://doi.org/10.1177/0748175611418522.

Bates, D., Mächler, M., Bolker, B. and Walker, S. (2015) 'Fitting linear mixed-effects models using lme4', *Journal of Statistical Software*, 67(1), pp. 1–48. Available at: https://doi.org/10.18637/jss.v067.i01.

Cheung, A.C.K. and Slavin, R.E. (2013) 'The effectiveness of educational technology applications for enhancing mathematics achievement in K-12 classrooms: a meta-analysis', *Educational Research Review*, 9, pp. 88–113. Available at: https://doi.org/10.1016/j.edurev.2013.01.001.

Cilesiz, S. and Greckhamer, T. (2020) 'Qualitative comparative analysis in education research: its current status and future potential', *Review of Research in Education*, 44(1), pp. 332–369. Available at: https://doi.org/10.3102/0091732X20907347.

DfE (2015) *National curriculum assessments at key stage 2 in England (revised)*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/483897/SFR47_2015_text.pdf (Accessed: 14 December 2025).

DfE (2022a) *Implementation of education technology in schools and colleges*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1107808/Implementation_of_education_technology_in_schools__and_colleges.pdf (Accessed: 1 October 2023).

DfE (2022b) *School-led tutoring guidance*. Available at: https://assets.publishing.service.gov.uk/media/62601734e90e072a014d517d/Updated_School-Led_Tutoring_Guidance_.pdf (Accessed: 5 March 2024).

DfE (2023) *Schools, pupils and their characteristics, Academic year 2022/23*, *GOV.UK*. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics (Accessed: 4 March 2024).

DfE (2025) *Key stage 2 attainment: academic year 2024/25*, *GOV.UK*. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2024-25 (Accessed: 12 November 2025).

EEF (2019) *Putting evidence to work: a school's guide to implementation*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/eef-guidance-reports/implementation/EEF_Implementation_Guidance_Report_2019.pdf?v=1696494696 (Accessed: 1 October 2023).

EEF (2022a) *Best evidence on impact of COVID-19 on pupil attainment*. Available at: https://educationendowmentfoundation.org.uk/guidance-for-teachers/covid-19-resources/best-evidence-on-impact-of-covid-19-on-pupil-attainment (Accessed: 6 March 2024).

EEF (2022b) *Statistical analysis guidance for EEF evaluations*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1709155078 (Accessed: 6 March 2024).

European Union (2016) *Regulation (EU) 2016/679 of the European Parliament and of the Council (General Data Protection Regulation)*. *Legislation.gov.uk*. Available at: https://www.legislation.gov.uk/eur/2016/679 (Accessed: 3 June 2026).

Fox, J., Kleiber, C., Zeileis, A., Kuschnig, N., and R Core Team (2026) 'ivreg: Instrumental-Variables Regression by "2SLS", "2SM", or "2SMM", with Diagnostics'. Available at: https://cran.r-project.org/web/packages/ivreg/index.html (Accessed: 3 June 2026).

Harrison, W. and Higgins, S. (2023) *A global evidence review for policymakers: findings from meta-analyses or tutoring*. Available at: https://www.edt.org/research-and-insights/a-global-evidence-review-for-policymakers-findings-from-meta-analyses-of-tutoring/#:~:text=02%2F08%2F2023-,A%20global%20evidence%20review%20for%20policymakers,from%20meta%2Danalyses%20of%20tutoring&text=Tutoring%20is%20widely%20acknowledged%20as,of%20the%20Covid%2D19%20pandemic. (Accessed: 6 October 2023).

Haßler, B., Huntington, B., Klune, C., Lester, J., Bhutoria, A. and Mansour, H. (2025) *Understanding quality characteristics of EdTech interventions and implementation for disadvantaged pupils*. Available at: https://doi.org/10.53832/opendeved.1170.

Higgins, S., Xiao, Z. and Katsipataki, M. (2012) *The impact of digital technology on learning: a summary for the Education Endowment Foundation*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/guidance/The_Impact_of_Digital_Technologies_on_Learning_2012.pdf?v=1765134312 (Accessed: 15 December 2025).

Hodge, L., Little, A. and Weldon, M. (2021) *GCSE attainment and lifetime earnings*. Available at: https://assets.publishing.service.gov.uk/media/60c36f0cd3bf7f4bd11a2326/GCSE_Attainment_and_Lifetime_Earnings_PDF3A.pdf (Accessed: 6 March 2024).

Jiang, Y. and Small, D. (2014) 'ivpack: Instrumental Variable Estimation (Version 1.2)'. Available at: http://download.nust.na/pub3/cran/web/packages/ivpack/ivpack.pdf (Accessed: 3 June 2026).

Lewin, C., Smith, A., Morris, S. and Craig, E. (2019) *Using digital technology to improve learning: evidence review*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/guidance/Using_Digital_Technology_to_Improve_learning_Evidence_Review.pdf?v=1765283244 (Accessed: 14 December 2025).

Li, Q. and Ma, X. (2010) 'A meta-analysis of the effects of computer technology on school students' mathematics learning', *Educational Psychology Review*, 22(3), pp. 215–243. Available at: https://doi.org/10.1007/s10648-010-9125-8.

Mavrikis, M., Schlepps, D.S. and Mubeen, J. (2018) 'Short-term efficacy and usage recommendations for a large-scale implementation of the Math-Whizz Tutor', in C. Penstein Rosé, R. Martínez-Maldonado, H.U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. Du Boulay (eds) *Artificial Intelligence in Education (AIED 2018)*.

Cham: Springer. Available at: http://link.springer.com/10.1007/978-3-319-93846-2_85 (Accessed: 29 September 2023).

Menzies, L., Ramaiah, B. and Boulton, C. (2021) *A space for maths: exploring the need for maths tutoring and the potential role of Third Space Learning*. Available at: https://cfey.org/wp-content/uploads/2021/09/A-Space-for-Maths-Sept-2021-for-publication-amended.pdf (Accessed: 6 March 2024).

Outhwaite, L., Early, E., Herodotou, C. and Van Herwegen, J. (2022) *Can maths apps add value to young children's learning?* Available at: https://www.nuffieldfoundation.org/wp-content/uploads/2022/05/Can-Maths-Apps-Add-Value-to-Young-Childrens-Learning-A-Systematic-Review-and-Content-Analysis_Web_final_v2.pdf (Accessed: 6 March 2024).

Renaissance Learning and Education Policy Institute (2021) *Understanding progress in the 2020/2021 academic year: interim findings*. Available at: https://assets.publishing.service.gov.uk/media/6239cbb1e90e07799a20bc0e/Understanding_progress_in_the_2020_21_academic_year_Interim_findings_Jan2021.pdf (Accessed: 6 March 2024).

Richardson, M., Golding, J., Isaacs, T., Barnes, I., Wilkinson, D., Swensson, C. and Maris, R. (2025) *TIMSS 2023: National report for England Volume 2*. London: DfE.

Sahasranaman, A., Aston, K., Schwendel, G. and Welbourne, S. (2025) *Evaluation of whizz education: maths-whizz intelligent tutoring programme. Evaluation protocol*. EEF. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/maths_whizz_intellingent_tutoring_-_evaluation_protocol.pdf?v=1762877165 (Accessed: 12 November 2025).

Schwendel, G., Welbourne, S. and Sahasranaman, A. (2025) *Evaluation of whizz education: maths-whizz intelligent tutoring programme. Statistical analysis plan*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/eemw_-_statistical_analysis_plan.pdf?v=1762877165 (Accessed: 13 November 2025).

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Higgins, S., Einbeck, J., Culliney, M. and Demack, S. (2023) *Improving power calculations in educational trials*. London: EEF. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/improving-power-calculations-in-educational-trials (Accessed: 2 October 2023).

Steenbergen-Hu, S. and Cooper, H. (2013) 'A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning', *Journal of Educational Psychology*, 105(4), pp. 970–987. Available at: https://doi.org/10.1037/a0032447.

Stringer, E., Lewin, C. and Coleman, R. (2019) *Using digital technology to improve learning: guidance report*. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/eef-guidance-reports/digital/EEF_Digital_Technology_Guidance_Report.pdf?v=1765355881 (Accessed: 14 December 2025).

The Sutton Trust (2024) ‘General election policy briefing: closing the attainment gap’. Available at: https://www.suttontrust.com/wp-content/uploads/2024/02/Closing-the-attainment-gap.pdf (Accessed: 10 February 2026).

Treadaway, M. (2017) *Long-term disadvantage, part three: ethnicity, EAL and long-term disadvantage*, *FFT Education Datalab*. Available at: https://ffteducationdatalab.org.uk/2017/07/long-term-disadvantage-part-three-ethnicity-eal-and-long-term-disadvantage/ (Accessed: 14 December 2025).

UK Government (2018) *Data Protection Act 2018*, *Legislation.gov.uk*. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents (Accessed: 3 June 2026).

Whizz Education (2021) *Whizz Education proof pack*. Available at: https://www.whizzeducation.com/wp-content/uploads/2022/02/2021-Proof-Pack-Whizz.pdf (Accessed: 6 March 2024).

Worth, J. and McLean, D. (2025) *Teacher labour market in England annual report 2025*. Available at: https://www.nfer.ac.uk/publications/teacher-labour-market-in-england-annual-report-2025/ (Accessed: 14 December 2025).

YouGov (2014) ‘National numeracy survey results’. Available at: http://cdn.yougov.com/cumulus_uploads/document/b6j344revv/YouGov-Survey-Results-National-Numeracy-140227.pdf (Accessed: 6 March 2024).

Appendix A: The EEF cost rating

Appendix A Figure 1: 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

Table 1: Padlock assignment

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%	X				X
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 **[X]**		
2	Design for comparison that considers selection only on some relevant confounders	0.40 - 0.49	31-40%					
1	Design for comparison that does not consider selection on any relevant confounders	0.50 - 0.59	41-50%					
0	No comparator	>=0.6	>50%					

Threats to validity	**Risk rating**	**Comments**
Threat 1: Confounding	**Low**	Well thought through and well conducted randomised design, good level of balance between the groups and appropriate controls included.
Threat 2: Concurrent Interventions	**Low/moderate**	Interleaved design reduces that risk as intervention and control are within the same school. Adequate provision to ensure no confounding with similar interventions.
Threat 3: Experimental effects	**Low**	Hawthorne and John Henry effects are unlikely with this design and implementation. There may be some risk of contamination, but this has been explored and while possible it remains unlikely with the information provided. Limited risk of contamination as control group pupils are in the same schools but not same year groups as intervention pupils. But control group pupils would not have access to Maths Whizz themselves, contamination could only be indirect via teachers and IPE suggests this was not an issue.
Threat 4: Implementation fidelity	**Low/moderate**	The study was an efficacy trial, and the IPE has reported school leaders were committed to and capable of implementing the adaptive tutoring. A high proportion of classes were compliant; however, it was set a lower level than the design suggested. Most teachers provided sufficient class time. Fidelity overall was good although schools did not always implement Maths Whizz as additional to other maths teaching time, or for as long as the developers recommend. If anything, these factors would tend to lead to under-estimates of the impact.

Threat 5: Missing Data	**Low**	Missing data is low to moderate, and while there is some differential missingness, it is minor. Analysis accounting for missing data shows similar results to that for complete cases
Threat 6: Measurement of Outcomes	**Low**	The evaluator chose an independently set-up and administered commercial test, provided the histograms which showed no issues around floor and ceiling effects. Appropriate choice of outcome measures and carried out by external assessors rather than school staff.
Threat 7: Selective reporting	**Low**	Reporting is very comprehensive. The evaluator has registered the study, followed the protocol and SAP, clearly highlighted where there were slight deviations, and has covered a wide range of subgroups. It is an extensive and well-presented analysis.

- **Initial padlock score:** [4] Padlocks – Well designed and well conducted RCT, low MDES, and approximately 13% missing data.
- **Reason for adjustment for threats to validity**: [0] Padlocks – All threats to validity were judged to be low.
- **Final padlock score:** initial score adjusted for threats to validity = [4] Padlocks
- **Reason(s) for final padlock rating:** No adjustments were considered necessary as all threats to validity were considered to be low.

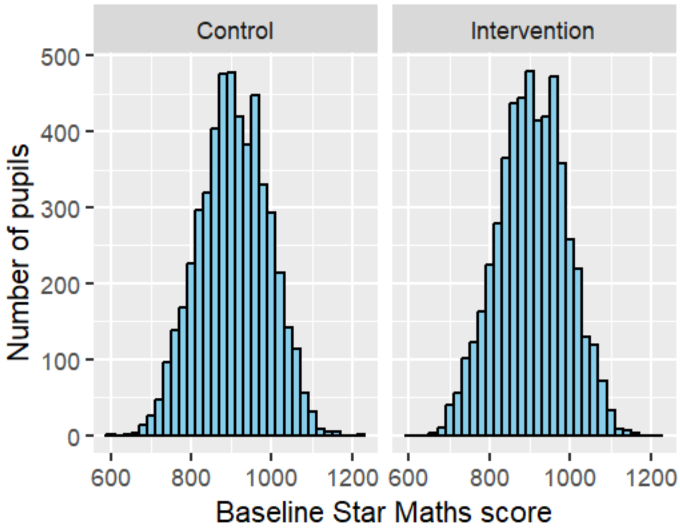
Appendix C: Effect size estimation

Appendix C Table 1: Effect size estimation

Outcome	Unadjusted differences in means	Variance components obtained from a model with no predictors						Pooled variance	
		Adjusted differences in means	N intervention (missing); N control (missing)	Between-school variance	Between-year group variance	Between-class variance	Within-class variance	Two groups	Eight groups
Star Maths Unified Scaled Score	6.24	5.25	4,568 (665); 4,415 (721)	0	3048.28	51.43	4396.45	7496.68	4468.85
Star Maths Age-Standardised Score	1.213	1.17	4,568 (665); 4,415 (721)	14.19	5.76	2.66	216.09	238.35	236.31
FSM subgroup – Star Maths Unified Scaled Score	9.82	9.49	1,176 (248); 1,125 (276)	0	2595.79	59.40	4822.65	7487.83	4549.70

Appendix D: Distribution of baseline Star Maths Unified Scaled Scores for control and intervention pupils

Appendix D Figure 1: Histogram of baseline Star Maths Unified Scaled Scores for control and intervention pupils



Appendix E: Maths and Me Survey

Evaluation of Whizz Education: Maths-Whizz Intelligent Tutoring Programme

Maths and Me

These questions ask how you feel about maths.

Please choose ONE answer for each question. Be sure to answer ALL the questions.

Remember that there are no "right" and "wrong" answers. These are about how *you* feel about maths.

Please say how you feel about each of the following statements. Please choose one answer per row.						
		Not at all like me	Not like me	Sort of like me	Like me	Very like me
1.1	I am really good at maths.					
1.2	I love maths.					
1.3	I understand maths.					
1.4	Maths is boring.					
1.5	I can solve difficult maths problems.					
1.6	I enjoy doing maths puzzles.					
1.7	Maths is very hard for me.					
1.8	I enjoy doing maths problems on my own "just for fun".					
1.9	Maths is confusing to me.					
1.10	Maths is fun.					
1.11	I look forward to learning new maths.					
1.12	Maths comes easily to me.					
1.13	I hate maths.					
1.14	I enjoy playing maths games.					
1.15	I can tell if my answers in maths make sense.					
1.16	I enjoy studying maths.					
1.17	Doing maths is easy for me.					
1.18	Solving maths problems is fun.					

Additional information about the Maths and Me instrument can be found here: https://doi.org/10.1177/0748175611418522.

Further appendices

Appendix F – J can be found published as a separate document ‘Further Appendices’ on the project page.

Education Endowment Foundation

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

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn