Education Endowment Foundation

WhatWorked Teachers Feasibility Report

Methodological Innovation report

November 2026

Dr Emma Dobson, Dr Wayne Harrison, Professor Steve Higgins, Rahil Khowaja, and Germaine Uwimpuhwe

What Worked Education

The Education Endowment Foundation is an independent charity dedicated to breaking the link between family income and education achievement. We support schools, nurseries, and colleges 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.** 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 re-endowed with an additional £137m, allowing us to continue our work until at least 2032.

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

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

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator 3

Executive summary 4

Introduction 6

Methods 13

Findings 23

Discussion 27

Conclusion 35

References 37

Appendices 40

About the evaluator

This project was designed and evaluated by the research team at WhatWorked Education: Dr Wayne Harrison, Professor Steve Higgins, Rahil Khowaja, Germaine Uwimpuhwe, and Dr Emma Dobson (Principal Investigator).

Contact details:
WhatWorked Education
Durham Workspace
Abbey Road
Pity Me
Durham
DH1 5JZ

emma.dobson@whatworked.education

Executive summary

The project

This project aimed to explore the feasibility of using *WhatWorked Teachers*, a web-based evaluation platform developed by WhatWorked Education, to conduct small-scale randomised controlled trials as part of teachers' routine classroom practice. The study involved teachers working with pupils in Years 2 to 5 in mainstream primary schools in England. Teachers selected a spelling approach, either a front-loaded Look, Say, Cover, Write, Check (LSCWC) approach (10 minutes of daily/weekly practice where all spellings are given to pupils at the start of the week) or an incremental LSCWC approach (a daily targeted 5-minute session is used to introduce new spelling words incrementally throughout the week), and a randomisation option (individual randomisation, where only some pupils would receive the approach or class randomisation, where whole classes would be randomised to implement the approach). The *WhatWorked Teachers* platform then conducted the randomisation allocation. Pupils were given a pre-test, then the spelling approaches were delivered to the randomly allocated pupils or classes over the course of six weeks; finally, a pupil post-test was taken. Teachers entered the pupil data onto the platform and were then provided with an instant impact report. The *WhatWorked Teachers* platform collected cumulative data from all the small-scale classroom trials.

The evaluation was conducted by WhatWorked Education, who also developed the platform. The evaluation was designed to explore the feasibility of: engaging schools and teachers; teachers using the platform to select trials; teachers adhering to the trial protocols and implementation of the approaches; and drawing conclusions from the data gathered. Across the two phases of recruitment and delivery, the final sample included 18 teachers and 532 pupils from 15 schools. The evaluation used platform usage data, teacher surveys, and teacher interviews. Phase 1 delivery took place in September and October 2025. Phase 2 delivery took place in November and December 2025.

Table 1: Summary of findings

Area of research	Findings
Recruitment and engagement	There were 58 schools that expressed an interest in the project, with 19 signing up to take part (33% conversion rate). Teacher accounts were created in 15 schools, and 42 teacher accounts were created, 29 of which were judged to be legitimate teacher accounts (not duplicates or senior managers). Then 25 teachers selected a trial, and 23 completed the randomisation stage. Teachers reported some uncertainty around the trial selection part of the process. The front-loaded spelling approach and class-level randomisation were the more popular options, chosen by 15/18 teachers (83%) and 10/18 (56%) respectively. Then 11/18 teachers completed the teacher surveys and 6 participated in case study interviews.
Trial implementation and attrition	Teachers who progressed to implementation (18/29, 62%) were able to deliver the trial procedures with all 18/18 (100%) of those who administered the pre-tests also completing the post-tests. Front-loaded trials randomised 420 pupils and retained post-test data for 413 pupils (98.3%), while incremental trials randomised 112 pupils and retained post-test data for 93 pupils (83.0%). Teacher confidence with using the platform was high. Platform guidance, assessment administration, and data uploads were reported to be manageable, and the six-week implementation period fitted well within a school half-term. Implementation was more consistent for whole-class trials than for small-group trials, where staffing pressures, timetabling, and maintaining separation between intervention and comparison groups presented greater challenges. The potential risk for contamination was also less likely for whole-class compared to small-group trials.
Interpretation of data	The platform successfully generated standardised statistical analyses and automated impact reports following submission of post-test data. Teachers had some difficulties with interpreting the statistical content in the impact reports, especially effect sizes, confidence intervals, and the meaning of statistical uncertainty. However, some teachers reported discussing findings with colleagues and senior leaders, suggesting potential to stimulate evidence-informed professional dialogue. Due to the different sample sizes in the trials of the front-loaded approach and the low

	number of trials of the incremental approach, the findings do not provide sufficient evidence to establish a typical effect size and should be interpreted as exploratory. The platform was able to conduct an exploratory meta-analysis, combining the data from the 15 front-loaded trials to explore overall patterns in the findings.

Additional findings

Recruiting schools to this project may have been more challenging than for typical EEF trials. The project was described to schools as a test of the feasibility of a new evaluation platform, rather than a project that was assessing the impact of a programme or approach on pupil attainment. This may have made the potential benefits of participation less immediately apparent. Teacher uncertainty around the trial selection at the outset of the platform process, alongside practice constraints including timetabling, staffing changes, and competition priorities, may have accounted for the attrition between account creation and implementation. Of the 29 teachers who created an account, 11 did not progress to implementing the trial (attrition rate of 38%). Future versions of the platform could simplify and improve the clarity of the onboarding process to support teacher selection of trial designs.

Once teachers had started the trials, retention was very high: all 18 teachers who began implementation completed their trials. This suggests teachers were able to deliver the trials easily within routine classroom practice. Teachers were generally able to enter pupil data, administer assessments, and submit outcome data onto the platform independently and were found to understand the rationale for random allocation. Survey and interview responses indicated positive experiences of using the platform, with teachers valuing the opportunity to test approaches within their own classrooms.

The imbalance between the number of teachers selecting the front-loaded spelling approach, compared with the incremental approach, meant that the impact estimates from the incremental approach had substantially wider confidence intervals indicating greater uncertainty around the estimated effects. Fewer teachers selected the individual randomisation option, and there was some indication from survey responses that those who had chosen this small group option of delivery found it more difficult to implement. These constraints should be considered when designing trials using the *WhatWorked Teachers* platform, but whilst small-group design may have greater implementation challenges, their use should not be precluded, as they can enable participation from small schools who cannot randomise whole classes and can allow the evaluation of targeted interventions.

The challenge with teachers' interpretation of the impact reports is a key design challenge for the *WhatWorked Teachers* platform. The presentation of results to teachers will need to be refined to improve accessibility for teachers and also to more clearly communicate the limitations of any findings from small classroom trials. Future development should focus on Plain English summaries, visual presentation of statistical results such as confidence intervals (CI), and explicit guidance regarding the appropriate use of any exploratory findings.

Meta-analysis of the cumulative data demonstrated that the platform was able to produce exploratory analysis of pooled findings. The study provides insights into the range of effect sizes that may arise in teacher-delivered classroom RCTs (randomised controlled trials), as well as their statistical limitations. For the front-loaded condition, the pooled effect size was 0.27 (95% CI: 0.03 to 0.51). This is larger than is typically observed in educational Randomised Controlled Trials (RCTs). However, despite pooling the data, the confidence interval remains wide, suggesting that there are still questions around the precision of the estimate. More work will be needed to examine appropriate statistical approaches for analysing cumulative classroom evaluation evidence and also appropriate safeguards for its interpretation.

Introduction

This study examined the feasibility of using the *WhatWorked Teachers* platform to support teacher-led, small-scale randomised controlled trials in primary school classrooms. Teachers used the platform to compare short spelling approaches derived from Look, Say, Cover, Write, Check (LSCWC). The core evaluative question, however, was not whether a particular spelling routine 'worked', but whether teachers could use the platform to implement a short, structured, and methodologically credible classroom micro-trial under everyday school conditions, and whether they could understand and appropriately interpret the automated impact reports generated by the platform to inform evidence-based decision making.

Background evidence

The *WhatWorked Teachers* platform has been developed to enable teachers to conduct rigorous, small-scale randomised controlled trials as part of routine classroom practice. Throughout this report, the term micro-randomised controlled trial (micro-RCT) refers to a small-scale classroom randomised controlled trial designed to evaluate a specific teaching practice, pedagogical routine, or component of a broader educational intervention within authentic school settings. In this context, the term *micro* refers to the scale and scope of the evaluation, characterised by implementation within individual classrooms or small groups of classes, relatively short trial durations, and focused evaluation of discrete teaching practices. It does not refer to repeated within-participant randomisation.

Micro-RCTs enable teachers to evaluate specific pedagogical practices under everyday classroom conditions whilst maintaining the methodological principles of experimental evaluation. By randomly allocating pupils (or classes) to alternative instructional approaches, teachers can generate robust evidence about what works within their own context while minimising bias associated with informal comparisons. The smaller scale of these evaluations makes them feasible to integrate into routine teaching and allows schools to undertake successive cycles of evidence generation as part of ongoing professional enquiry. Although individual micro-RCTs are not intended to provide definitive estimates of intervention effectiveness in isolation and will typically have limited statistical power, evidence generated across multiple comparable trials can subsequently be synthesised using cumulative meta-analysis to provide more precise estimates of intervention effectiveness.

The educational use of the term *micro-RCT* should be distinguished from the concept of Micro-Randomised Trials (MRTs) described within behavioural medicine (Liao *et al.*, 2016). In healthcare research, MRTs involve repeated sequential randomisation of the same participant across numerous decision points to optimise adaptive interventions. Although both approaches share an interest in iterative experimentation and evidence generation, they address different research questions and employ different randomisation strategies. The present study adopts a conventional parallel-group randomised controlled trial design implemented at a smaller scale within routine classroom practice. The term *micro-RCT* is therefore used throughout this report in accordance with the educational tradition of small-scale classroom experimentation rather than the repeated-randomisation methodology used in behavioural medicine.

Small-scale experimental enquiry has a long history within educational research. Professor Carol Fitz-Gibbon was an early advocate of classroom-based 'micro-enquiry', encouraging teachers to undertake randomised comparisons of alternative teaching approaches within their own classrooms to generate evidence that could directly inform professional practice (for example, what time of day is best for maths?). This philosophy was subsequently developed through the National College for Teaching and Leadership's 'Closing the Gap: Test and Learn' programme (Churches, 2016), which demonstrated that teachers could successfully conduct both larger-scale school randomised trials and smaller classroom-based randomised evaluations as part of professional learning. Churches (2016) argues that 'through micro-enquiry teachers can take control, take responsibility; it empowers them to make more complex and sophisticated professional decisions within their own professional lives' (p. 62). This approach aligns closely with the principles of evidence-informed practice and participatory action research by positioning teachers as active producers, rather than passive consumers, of research evidence.

Despite this conceptual foundation, relatively little empirical research has examined the feasibility of teachers independently implementing classroom randomised evaluations at scale, or of supporting this process through dedicated

digital platforms. A rapid scoping search identified relatively few published examples of classroom micro-RCTs (for example, Mouton *et al*., 2018; Topping *et al*., 2012). Although limited in number, these studies suggest that small-scale classroom experiments can support professional learning whilst generating evidence that is directly relevant to local educational contexts and contribute to a broader understanding of effective interventions through the aggregation of individual trials via cumulative meta-analysis. However, they also highlight the practical challenges of implementing rigorous experimental methods within busy school environments, including trial design, randomisation, data management, and statistical analysis. The *WhatWorked Teachers* platform was developed to address these implementation challenges by automating the technical aspects of trial delivery whilst enabling teachers to remain the primary implementers of the evaluation. Rather than requiring specialist research expertise, the platform guides teachers through a structured evaluation process, automates participant randomisation, supports data collection, performs statistical analyses, and generates impact reports to aid interpretation of trial results.

The aptitude and ability of teachers to conduct randomised controlled trials in schools in England has been previously explored in the 2016 *Closing the Gap: Test and Learn* research report by the National College for Teaching and Leadership (Churches, 2016), which recorded the experiences of teachers participating in RCTs as part of professional development initiatives aimed at improving teaching practices and student outcomes. Contrary to assumptions made at the start of the programme, the report found that teachers were not resistant to the use of control groups, engagement in statistical research, or to the use of RCTs.

'Probably the most important finding is that there is capacity within the system for teachers to do rigorous, small-scale enquiries that identify whether or not approach A is better than approach B. That is really empowering—if we are going to move towards a school-led system of school improvement, that part of the ecology needs to be there. It still needs larger-scale studies and other kinds of research, but schools definitely need micro-level enquiry to test what they think is going to be effective and make sure it actually is.'

(Churches, 2016, **p. 62**)

Many participants noted that their engagement in RCTs made them more reflective practitioners, broadened their perspectives, and challenged their assumptions about teaching and learning. One participant from the report emphasised this impact, stating, 'Research thinking is what you normally do in the classroom. Actually being engaged in the formalised process has brought into conscious awareness what outstanding teachers do by default' (Churches, 2016, **p. 59**). The report concluded that teachers have the capacity to undertake rigorous small-scale experimental enquiry and argued that such approaches have an important role within a self-improving, evidence-informed education system.

Potential advantages and implementation challenges of classroom micro-RCTs

If they can be implemented successfully, classroom micro-RCTs have the potential to offer several advantages over traditional approaches to educational evaluation. Their smaller scale and integration within routine teaching may enable teachers to generate rigorous evidence about specific pedagogical practices while reducing many of the practical barriers associated with conventional large-scale randomised controlled trials. Potential advantages include:

1. Rapid evaluation within routine practice: Classroom micro-RCTs can be conducted over relatively short periods, enabling teachers to evaluate discrete teaching practices or components of broader interventions within everyday classroom settings. This offers the potential for schools to generate timely evidence that can inform ongoing teaching and learning.

2. Reduce implementation burden: By embedding evaluations within existing classroom practice and making use of routinely collected formative assessment data, classroom micro-RCTs may reduce many of the financial, logistical, and organisational demands typically associated with externally managed effectiveness trials.

3. Teacher engagement with research: Classroom micro-RCTs position teachers as active implementers of educational evaluation rather than passive recipients of research findings. This has the potential to strengthen research literacy, promote reflective practice, and support evidence-informed professional development by enabling teachers to evaluate approaches within their own classrooms.

4. Successive cycles of evidence generation: The relatively short duration of classroom micro-RCTs means that schools may be able to undertake multiple evaluations over time, supporting continuous professional enquiry and iterative refinement of teaching practice.

5. Contribution to a cumulative evidence base: Whilst individual classroom micro-RCTs are unlikely to provide definitive estimates of intervention effectiveness, multiple comparable trials conducted across different classrooms and schools may be synthesised using cumulative meta-analysis. If implemented consistently and at sufficient scale, this approach has the potential to generate increasingly precise estimates of intervention effectiveness whilst retaining the contextual relevance of classroom-based research.

These potential advantages should, however, be considered alongside important methodological limitations. Individual classroom micro-RCTs typically involve relatively small numbers of pupils or classes and therefore have limited statistical power. As a consequence, estimates of intervention effects are likely to be associated with wide confidence intervals and increased risks of both Type I (false positive) and Type II (false negative) errors. Individual trial findings should therefore be interpreted cautiously and should not be regarded as definitive evidence that an intervention is effective or ineffective within a particular setting. Rather, the primary value of classroom micro-RCTs lies in supporting structured local enquiry whilst contributing comparable evidence that can subsequently be synthesised across multiple trials. Whether this model can generate sufficiently robust evidence at scale remains an important empirical question and forms part of the rationale for the present feasibility study.

From a policy perspective, classroom micro-RCTs also have the potential to complement existing approaches to educational evaluation. If implemented at sufficient scale and with appropriate methodological consistency, evidence generated across multiple schools could be aggregated through cumulative meta-analysis to support understanding of intervention effectiveness across a range of educational contexts. As Churches (2016) argues: ‘toolkit summaries of research are effective to some degree, but these need to be tested at a micro level. Such an approach, if co-ordinated nationally, could build an evidence base to support implementation and understand variation related to context’ (**p. 72**). However, achieving this ambition depends upon sufficient numbers of comparable trials, standardised outcome measures, consistent implementation, and robust quality assurance procedures. Platforms such as *WhatWorked Teachers* have the potential to facilitate this process by standardising trial procedures and automating key methodological components, although whether this can be achieved in practice remains to be established.

Alongside these potential advantages, classroom micro-RCTs present several practical and methodological challenges that warrant careful consideration.

- Successful delivery depends upon teachers being willing and able to integrate evaluation activities into existing classroom practice. Competing workload demands, varying levels of research confidence and maintaining engagement throughout the evaluation cycle may all influence implementation fidelity and trial completion.

- Teacher-led evaluations inevitably provide less external control than researcher-managed trials. Variation in intervention delivery, assessment administration and data entry may introduce additional sources of measurement error or implementation bias. Digital platforms can help reduce some of these risks through standardised procedures and automated workflows, but cannot eliminate them entirely.

- Classroom evaluations are generally limited to outcome measures that are feasible to administer within routine school practice. While this enhances ecological validity, it may restrict the breadth or sensitivity of measures available compared with more resource-intensive research studies.

- Classroom micro-RCTs are typically small, meaning that individual effect estimates are subject to considerable statistical uncertainty. This places particular importance on ensuring that teachers are supported to interpret

findings appropriately and understand the limitations of individual trial results. Automated reporting should therefore communicate both estimated effects and the associated uncertainty, reducing the risk that decisions are based solely on the direction or magnitude of observed differences.

- Participation may be influenced by school capacity, leadership support, or teacher motivation, meaning that participating schools may not be representative of the wider education system. Furthermore, findings from individual classroom trials are necessarily context-dependent and should be interpreted alongside wider evidence rather than viewed in isolation.

Many of these challenges are not unique to classroom micro-RCTs but are common to educational randomised controlled trials more generally. However, they may become more pronounced within smaller, teacher-led evaluations because of reduced sample sizes, greater reliance on routine classroom implementation, and less centralised oversight. Consequently, before classroom micro-RCTs can be recommended as a scalable approach to educational evaluation, it is necessary to establish whether teachers can successfully implement them, whether the evidence generated can be interpreted appropriately, and whether a digital platform can provide sufficient methodological support to deliver robust and standardised evaluations under routine school conditions. These questions provide the central rationale for the present feasibility study.

Rationale for evaluation

Randomised controlled trials have become increasingly important within educational research and have substantially strengthened the evidence available to inform policy and practice. However, the vast majority of these evaluations are designed, implemented, and managed by researchers or external organisations, with schools acting primarily as sites for data collection rather than active participants in the research process. As Churches (2016, pp. 13–14) observes, 'Although a development to be applauded, the sudden expansion of trial evidence presents a potential problem… the accumulation of evidence is not going to be sufficient in itself to transform the use of such research without the direct involvement of practitioners'. Classroom micro-RCTs have the potential to address this challenge by supporting greater teacher involvement in educational evaluation. Rather than replacing large-scale effectiveness trials, they are intended to complement them by enabling teachers to evaluate specific pedagogical practices within their own classrooms using structured experimental methods. In doing so, they may help bridge the gap between externally generated evidence and classroom implementation, allowing teachers to examine how approaches identified through larger studies operate within their own educational contexts. As Churches (2016, p. 63) argues, 'Teacher-led RCTs have much potential as a means of mediating and exploring the effects of prior large-scale studies in the teacher's own context so that external solutions can be appraised according to local circumstances and priorities for spending'.

However, classroom micro-RCTs also present important methodological challenges. Because individual evaluations are typically conducted with relatively small numbers of pupils or classes, they are unlikely to provide sufficiently precise estimates of intervention effectiveness in isolation. Consequently, the primary purpose of individual classroom micro-RCTs is not to provide definitive evidence that a particular approach is effective or ineffective within a single school. Instead, their value lies in supporting structured local enquiry whilst generating standardised evidence that may subsequently contribute to a broader cumulative evidence base when synthesised across multiple comparable trials. Whether such an approach can produce reliable evidence at scale depends upon consistent implementation, standardised methodologies and sufficient numbers of comparable evaluations.

The *WhatWorked Teachers* platform has been developed to address many of the practical barriers that currently limit teacher participation in rigorous educational research. The platform automates key methodological components of classroom randomised trials including participant randomisation, data management, statistical analysis, and reporting, while enabling teachers to remain the primary implementers of the evaluation within their own classrooms. By reducing the technical expertise required to conduct experimental evaluations, the platform has the potential to make rigorous

classroom research more accessible to a wider range of schools, including those that may previously have lacked the capacity or resources to participate in externally managed research.

Given both the potential opportunities and the methodological uncertainties associated with this approach, a feasibility study is an essential first step before wider implementation can be recommended. The purpose of the present study is therefore not to establish the educational effectiveness of the spelling approaches included for evaluation within the platform, nor to demonstrate that individual classroom micro-RCTs can provide definitive evidence of effectiveness. Rather, it is to determine whether teachers can successfully use the *WhatWorked Teachers* platform to implement structured classroom micro-RCTs, whether the platform can support sufficiently robust and standardised trial procedures under routine school conditions, and whether teachers can appropriately interpret the evidence generated. Establishing these aspects of feasibility is a necessary prerequisite before evaluating the platform at a larger scale or investigating its potential contribution to a cumulative programme of educational evidence generation.

Evaluation objectives

The primary objective of this study was to assess the feasibility of implementing teacher-led micro-RCTs via an online platform as an approach to evaluate educational practices. Specific objectives included:

1. To evaluate the practicality of engaging the EEF research schools and their networks in conducting micro-RCTs via the *WhatWorked Teachers* platform.
2. To assess teacher participation, engagement, and compliance with micro-RCT protocols, leveraging implementation evaluation surveys and interviews.
3. To assess the feasibility of the platform from a methodological perspective, including understanding expected effect sizes.

As part of utilising the platform to run micro-RCTs, teachers input pre- and post-test assessment data. This data was used to explore secondary objectives to understand the typical effect sizes to be expected when micro-RCTs are utilised in this context and whether there is potential for micro-RCTs to produce cumulative evidence through meta-analysis.

Research questions

Primary research questions

1. To what extent can teacher-led micro-RCTs serve as a practical, robust method for evaluating educational practices?
    - Is it feasible to engage schools and teachers and get them to participate? What are the facilitators and barriers?
    - Is it feasible to get teachers to select a trial of interest, comply/adhere to protocols and implement the intervention? What are the facilitators and barriers?
    - Is it feasible to draw conclusions from the data gathered (including completeness, reliability, effect sizes etc.)? What are the challenges and can they be mitigated?

Secondary research questions

1. What are the typical effect sizes to be expected when micro-RCTs are utilised in this educational context?
2. Does the use of micro-RCTs have the potential to produce cumulative evidence through meta-analysis?

Ethical review

To ensure robust ethical oversight, this study was reviewed by WhatWorked Education during an internal ethics meeting at the design stage. This study protocol also underwent peer review by two external reviewers: one appointed by the EEF, and an independent academic reviewer, Professor Peter Tymms, invited by WhatWorked Education.

The ethical review considered the potential risks associated with the interpretation of evidence generated through classroom micro-RCTs. Because individual classroom trials typically involve relatively small sample sizes, effect estimates may be associated with substantial statistical uncertainty. To mitigate the risk of over-interpreting individual trial findings, the *WhatWorked Teachers* platform impact report (Appendix F) provides teachers with explanatory guidance describing how the statistical analysis was conducted, what effect sizes represent, and the limitations of interpreting results from small-scale evaluations. Teachers are explicitly reminded that findings from individual classroom trials should not be compared directly with larger national studies and are encouraged to use professional judgement alongside the statistical outputs. As the effectiveness of these supports had not previously been evaluated, assessing whether teachers could appropriately understand and interpret the automated reports formed one of the objectives of the present feasibility study.

The study prioritised the ethical considerations relating to participating pupils, teachers, and schools by ensuring that all trial activities aligned with routine teaching and learning practices already established within schools. This approach minimised potential risks to pupils, avoided significant disruption to timetabling and limited additional workload for participating teachers. As the interventions reflected normal educational activities and teachers acted as evaluators within their own classrooms, the evaluation was considered to represent low-risk educational research. Consent for participation in the evaluation was obtained from the headteacher of each participating school. Headteacher consent was considered appropriate because the study involved evaluation of routine classroom practice rather than activities extending beyond normal educational provision. Schools nevertheless retained the autonomy to implement additional parental information or consent procedures should they consider these appropriate within their local context. Headteacher consent was adopted for the following reasons:

1. Alignment with usual practice: The interventions evaluated were pedagogical activities that schools may already undertake as part of routine teaching and learning. The study therefore posed minimal additional risk to pupils, with the primary purpose being to support schools in evaluating the effectiveness of routine classroom practices using structured experimental methods.
2. Genuine uncertainty (equipoise): Participating teachers evaluated a single spelling approach against their existing business-as-usual spelling instruction. At the outset of the study, there was genuine uncertainty regarding whether either spelling approach would produce better outcomes than established classroom practice within participating schools. This uncertainty provided the ethical justification for random allocation to intervention and business-as-usual conditions while ensuring that no pupils were knowingly disadvantaged through participation.
3. Minimising burden: Requiring individual parental consent for low-risk evaluations of routine educational practice can create substantial administrative demands for schools. Schools retained the autonomy to implement their own consent procedures if they wished to do so or where required by local policy or circumstance.
4. Equity of participation: Evidence suggests that requiring individual parental consent can reduce participation rates among disadvantaged pupils (Shaw *et al*., 2015; Liu *et al*., 2017; Bonell *et al.*, 2023), introducing selection bias and potentially limiting the representativeness and inclusivity of educational research. Adopting headteacher consent for this low-risk evaluation supported more equitable participation while remaining proportionate to the nature of the study.

Data protection

Student data was fully anonymised. No personal information was collected beyond anonymised pre- and post-test assessment scores and eligibility for Pupil Premium funding. The study complied with GDPR (General Data Protection Regulation) and processed data under the legal basis of public interest. Participating schools signed a data sharing notice (see Appendix C) in addition to a Memorandum of Understanding (MoU) and were provided with links to WhatWorked Education's Privacy Notice and the Education Endowment Foundation's Privacy Notice. The processing of anonymised data was necessary to assess the feasibility and potential impact of the micro-practices tested.

Teacher IPE (Implementation Process Evaluation) responses were anonymised, and any identifying information about participating schools or individuals was removed. Limited personal data was collected from teachers (name, school, and work email address). The legal basis for this processing under the UK GDPR and the Data Protection Act 2018 was Article 6(1)(e)—processing necessary for the performance of a task carried out in the public interest. Teachers were provided with

information about the research through three short instructional videos explaining the platform, the intervention, and the trial design. They were also given access to supporting materials including the trial protocol and implementation guidance. These procedures ensured transparency, limited data collection to what was necessary for the research, and protected teachers' rights under GDPR.

Project team

This project was designed and evaluated by the research team at WhatWorked Education, who developed and delivered the *WhatWorked Teachers* platform. The team was responsible for study design, recruitment, trial implementation, data collection, analysis, and reporting.

Methods

Study design

The study evaluated the feasibility of using the *WhatWorked Teachers* platform to support teacher-led micro-randomised controlled trials in primary school classrooms using a mixed-methods feasibility design. Participating teachers implemented short trials comparing two variations of a spelling practice derived from the Look, Say, Cover, Write, Check (LSCWC) approach. The spelling approaches used in the pilot functioned as test cases for the micro-RCT method.

Spelling approaches
'Look, Say, Cover, Write, Check' is a widely used approach focusing on structured repetition and recall. LSCWC is a method of learning spelling, where children:

- study a word;
- say the word aloud clearly, paying attention to the spelling;
- cover it up with a piece of paper or their hand (a sheet with a fold down it makes this easier);
- write the spelling next to the original word;
- check their version to see if they got it right; and
- write the correct spelling again if necessary.

Recent evidence indicates that LSCWC is among the most commonly used instructional strategies in classroom practice in English primary schools (Esposito, Herbert, and Sumner, 2022). While intervention studies suggest that the approach can support improvement in spelling of taught words (for example, Fisher, Cozens, and Greive, 2007; Dymock and Nicholson, 2017), its effectiveness relative to alternative approaches remains debated. Its widespread use in routine practice makes it an appropriate candidate for feasibility testing within teacher-led trials.

Spelling proficiency is a critical component of literacy development, directly influencing reading comprehension, vocabulary acquisition, and overall academic achievement (Graham and Santangelo, 2014). Despite its importance, recent reports indicate that a significant number of primary school students in England are struggling with spelling. In the 2021/2022 academic year, only 72% of pupils at the end of Key Stage 2 met the expected standard in grammar, punctuation, and spelling, a decline from 78% in 2018/2019 (DfE, 2022). In 2022/2023 and 2023/2024, 72% of pupils met the expected standard, unchanged from 2022 (DfE, 2024a; 2024b). Additionally, a 2024 Ofsted report highlighted that while reading instruction has improved, there is a need for greater focus on writing and spoken language skills, which encompass spelling abilities (Ofsted, 2024). These findings suggest that spelling remains a challenge for a considerable portion of primary school students in England, indicating a need for targeted interventions to enhance spelling proficiency.

Teacher-led micro-RCTs offer an innovative approach to identifying effective spelling approaches tailored to specific classroom contexts, potentially bridging gaps in attainment and reducing inequality. The focus of the LSCWC approaches offered to teachers was on improving spelling for primary school students in Year 2 to Year 5 (6–10 years of age). Variation 1 (front-loaded LSCWC) is ten minutes of daily/weekly LSCWC practice where all spellings are given to pupils at the start of the week. Each week, learners are introduced to ten new words and complete a practice formative test at the start of each session to review previous learning. Variation 2 (incremental LSCWC) is a working memory variation of LSCWC where a daily targeted 5-minute session is used to introduce new words incrementally throughout the week, rather than in a single session at the start of the week, helping to maximise working memory by reducing cognitive load and reinforcing retention through spaced repetition. Over the course of each week, learners are introduced to ten new words and then complete a practice formative test the following week to review previous learning. Teachers could select the target words they wished to use over the course of the intervention; these could be taken from the National Curriculum word lists for English spelling or could be derived from their own curriculum. Control conditions undertook business as usual, with no restrictions placed

on activity permitted as part of usual practice. Pupils in the control group could receive the approach in the following term if the approach was successful.

WhatWorked Teachers

These approaches were tested via 16 unique micro-RCT designs listed on the *WhatWorked Teachers* platform over six weeks. Alongside selecting one of the two spelling approaches, participating teachers could choose from two options for randomisation (individual within-class randomisation to test a targeted strategy delivered to small groups by a teacher or teaching assistant; or class randomisation to test a whole-class approach) and four year-group options (Year 2, Year 3, Year 4, and Year 5). If a teacher selected individual randomisation, the approach was delivered within a class. As a result, not all pupils within the class would receive the approach at the same time. Teachers used the *WhatWorked Teachers* platform to randomly allocate learners to intervention and control groups. Teaching assistants or teachers worked with small groups of students for ten minutes to practise using the spelling strategy and set it as homework. If a teacher selected class randomisation, the approach was delivered as a whole-class approach. All pupils within the class received the approach. Teachers use the *WhatWorked Teachers* platform to randomly allocate two classes per group to either intervention or control groups. Multiple trials could be conducted within a school across different year groups. Table 2 provides an overview of each micro-RCT design.

Table 2: Micro-RCT designs that could be selected by teachers to evaluate LSCWC spelling approaches

Approach: Look, Say, Cover, Write, Check (LSCWC)															
LSCWC Approach 1								LSCWC Approach 2							
Individual Randomisation				*Class Randomisation*				*Individual Randomisation*				*Class Randomisation*			
Year 2 Trial	Year 3 Trial	Year 4 Trial	Year 5 Trial	Year 2 Trial	Year 3 Trial	Year 4 Trial	Year 5 Trial	Year 2 Trial	Year 3 Trial	Year 4 Trial	Year 5 Trial	Year 2 Trial	Year 3 Trial	Year 4 Trial	Year 5 Trial
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#13	#14	#15	#16

Upon selecting a trial design, teachers viewed three short instructional videos: one introducing the platform and its functionalities, another explaining the approach and its implementation, and a third outlining the research design. Teachers also received a resource folder containing a trial protocol and relevant materials to ensure clear understanding of the research design, resource implications, and ethical considerations. Teachers who chose to proceed were guided through the randomisation and data entry process by the platform's Trial Wizard. Upon completion of each micro-trial, an automated ANCOVA is used to produce a teacher impact report (Appendix F). The ANCOVA uses pupil performance on a spelling test (marked out of 20) as a proximal measure of the impact of the intervention. Intervention and control pupils complete the same pre- and post-test assessment, allowing the analysis to adjust for baseline attainment. The assessments were administered by teachers as part of the classroom trial procedures, and results were entered into the platform. Teachers administered the pre-test before the intervention began and the post-test at the end of the trial period.

WhatWorked Teachers user flow

- Teachers select an approach of the LSCWC intervention they would like to evaluate (i.e., front-loaded or incremental).
- Teachers watch three guidance videos: 1) Welcome (what are micro-RCTs, who are WhatWorked Education), 2) Intervention (explanation of intervention theory of change), 3) Implementation (how to set up the intervention trial, deliver and assess the intervention, upload test scores to the platform, and access the impact report).
- Teachers download the resource folder (including teacher guide and FAQs (Appendix E), anonymous pupil test score spreadsheet, and example LSCWC grid) and trial protocol for their chosen intervention.
- Teachers enter the number of classes or individual pupils they need to randomise, and the *WhatWorked Teachers* platform randomises allocation to intervention or control groups.

- Teachers conduct pre-tests with intervention and control pupils and enter pre-test scores on the *WhatWorked Teachers* platform.
- Teachers deliver intervention sessions across the trial period (six weeks) whilst control pupils receive business-as-usual provision.
- Teachers conduct post-tests with intervention and control pupils and enter post-test scores on the *WhatWorked Teachers* platform.
- The *WhatWorked Teachers* platform calculates an ANCOVA and percentage difference result which is embedded in an impact report which teachers receive instantly upon upload of post-test results.

Recruitment

To recruit schools for the study, the EEF invited any mainstream primary school in England to participate through its website and shared information about the project on their social media platforms. The WhatWorked Education team also displayed information about the project on their website (Appendix C) and social media feeds with an invitation to contact them. The WhatWorked Education team has strong relations with the schools in their region and also distributed personal invitations to schools. Teachers could read a setting information sheet (Appendix A) before completing an Expression of Interest (EoI) form (Appendix B) and receiving a Memorandum of Understanding (MoU) and Data Sharing Agreement to complete (Appendix D).

Inclusion/exclusion criteria

Micro-RCT participants were primary school pupils taught by participating teachers in Year 2, Year 3, Year 4, or Year 5 classes. These year groups were selected as these were identified in teacher feedback as the most feasible classes (avoiding transition year groups such as Year 1 and Year 6, avoiding SATs preparation). Pre-test scores were not used as inclusion criteria. To take part, teachers had to:

- teach in a mainstream primary school in England;
- have at least one class in Years 2, 3, 4, or 5;
- be able to deliver the intervention and enter test scores between September and October 2025, respond to surveys, and attend an interview/focus group between November and December 2025; and
- not be engaged in other EEF literacy trials over the same period with the same year groups.

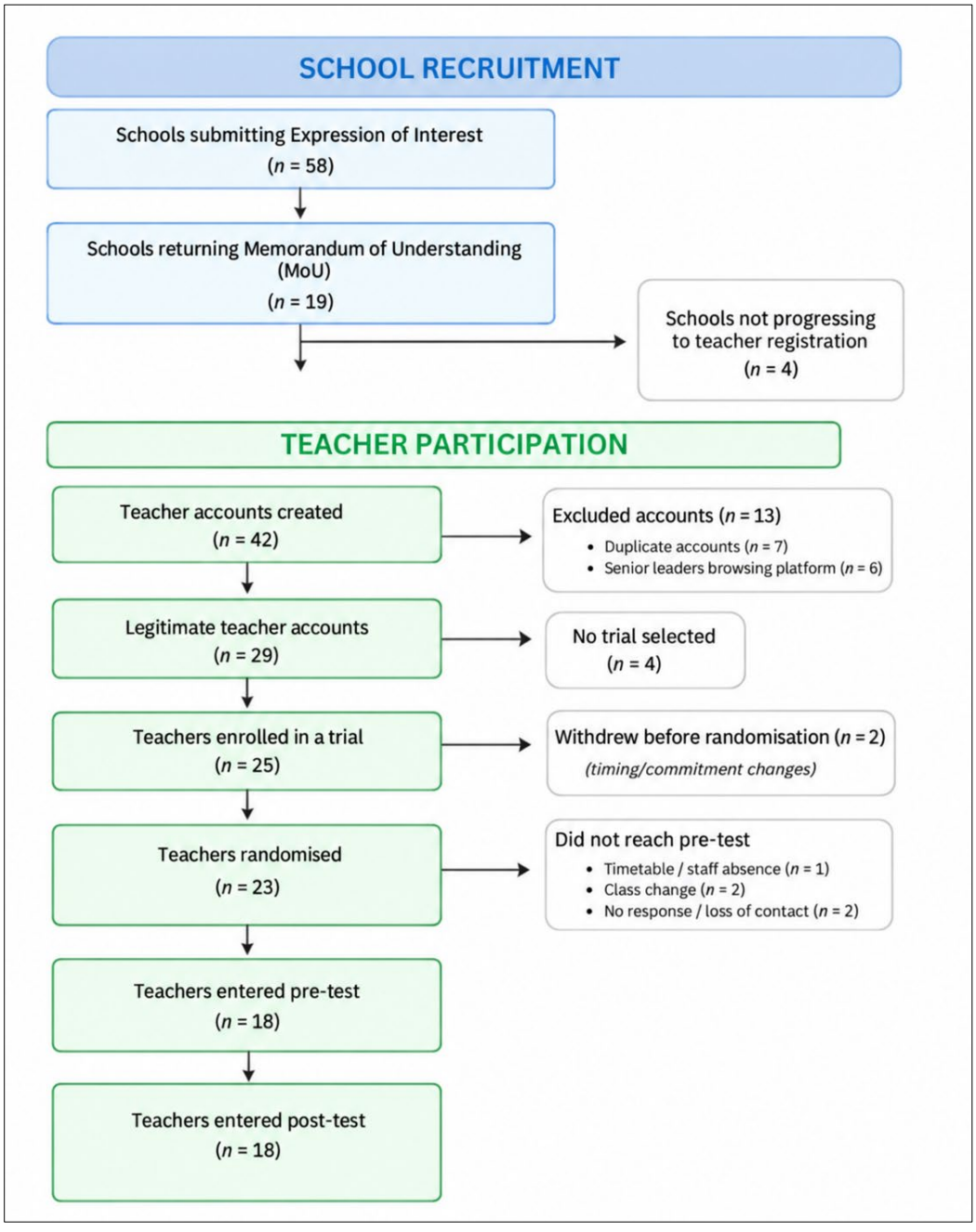
Recruitment process

The study aimed to recruit a minimum of 12 mainstream primary schools, 24 teachers, and approximately 500 pupils. An a priori power analysis was conducted for a three-level clustered design, with pupils nested within classes and classes nested within schools, following Hedges and Rhoads (2010). The calculation assumed a standardised effect size of 0.50, 80% power, a two-sided alpha of 0.05, school- and class-level ICCs of 0.20 and 0.10 respectively, and treatment-effect heterogeneity parameters of 0.50 at both school and class level. Assuming two participating classes per school, the calculation indicated that approximately 12 schools per allocation group would be required for class-level randomisation with 30 pupils per class, and 11 schools per allocation group for within-class randomisation with 16 pupils per condition. Full assumptions and formulae are provided in Appendix G. The feasibility study was not powered to detect effects within individual teacher-level trials, trial variants, year groups, or FSM (free school meals) subgroups; these analyses were exploratory.

Recruitment was undertaken in two phases to maximise participation and ensure sufficient teacher-level engagement to evaluate the platform under routine school conditions. The first recruitment phase took place between 28 April and 18 July 2025. This period coincided with one of the busiest phases of the primary school calendar, during which schools were managing statutory Key Stage 2 assessments, end-of-year reporting, transition activities, staffing arrangements for the following academic year, and preparation for the summer holidays. These competing priorities were anticipated to reduce schools' capacity to engage with new research initiatives and were considered likely to affect both recruitment and the speed with which expressions of interest progressed to formal participation.

Recruitment was undertaken initially at school level. Schools registered their interest by submitting an EoI, after which they were issued with an MoU for headteacher approval. Importantly, the MoU represented a school-level agreement to participate rather than confirmation of teacher numbers. Once an MoU had been returned, any teacher within the school could create a *WhatWorked Teachers* account and undertake a trial at the start of the new school year in September. Consequently, EoIs and MoUs should be interpreted as measures of school recruitment rather than teacher recruitment. Figure 1 summarises the flow of schools and teachers through each stage of recruitment.

Figure 1: Trial participation CONSORT diagram



It was not possible to determine from either Expressions of Interest or completed MoUs how many teachers would ultimately participate, as some schools contributed a single teacher while others contributed multiple teachers. During Phase 1, 43 schools submitted an EoI and 15 schools returned a signed MoU. Given the uncertainty regarding subsequent teacher participation, together with the anticipated challenges associated with recruiting during the latter part of the summer term for a September start, a second recruitment phase was undertaken between 2 September and 24 October 2025. The September start date for the study presented its own recruitment challenges. The first half-term is typically characterised by substantial organisational activity as schools establish new classes, finalise pupil groupings, implement

staffing changes, and settle pupils into new routines. Teachers may also experience timetable changes, pupil movement between classes, staff and pupil absence, and adjustments to teaching responsibilities during the opening weeks of term. Consequently, it introduced further uncertainty regarding which teachers and classes would ultimately be available to participate once teaching arrangements had stabilised.

Across both recruitment phases, 58 schools submitted an EoI and were issued with an MoU. Nineteen schools returned signed agreements, representing an EoI-to-MoU conversion rate of 33% (19/58). Although this conversion rate was lower than anticipated, it should be interpreted in light of both recruitment windows and the fact that the MoU secured institutional participation rather than individual teacher recruitment.

Of the 19 schools returning an MoU, 15 schools progressed to active participation by registering at least one teacher on the *WhatWorked Teachers* platform. Four schools did not progress beyond institutional agreement because no teachers subsequently registered to undertake a trial. Each participating school received a unique access code allowing any teacher within the school to access the platform's dedicated space for the evaluations set-up and made available for the purpose of this feasibility study, together with a trial set-up checklist to support implementation (Appendix E).

Teacher participation was measured independently through platform registration data. There were 42 teacher accounts created across the 15 participating schools. Following routine data cleaning, 13 accounts were identified as duplicate registrations or senior leaders accessing the platform to explore available evaluations rather than participate directly. This resulted in 29 legitimate teacher accounts created on the *WhatWorked Teachers* platform. Of these, four teachers created an account but did not select an evaluation to trial, leaving 25 teachers actively enrolled in the study. Then, 23 teachers progressed to the randomisation stage, representing 96% of the predefined teacher recruitment target (23/24). Five of these teachers subsequently withdrew before entering pre-test data because of timetable changes, staffing changes, or loss of contact. Consequently, 18 teachers entered pre-test data, and all 18 completed the full pre- to post-test trial cycle, representing 100% retention once trial implementation had commenced. At pupil level, 532 pupils were randomised across the front-loaded and incremental LSCWC evaluations, with post-test data available for 506 pupils, indicating high levels of pupil retention throughout the study.

Table 3 provides a detailed breakdown of the 15 participating schools, including school characteristics, recruitment phase, teacher participation, trial selected, sample size, and completion status. Most participating schools were urban (n = 12), with three rural schools represented. Nationally, around one-quarter of primary schools are in rural areas (DEFRA, 2025), so rural schools are somewhat underrepresented in the sample. This may reflect practical constraints of the trial design, as rural schools are more likely to have smaller or mixed-year-group classes, making randomisation at the class or group level more difficult. School size varied considerably, ranging from 125 to 451 pupils, reflecting a mix of smaller and larger primary settings. Levels of disadvantage also varied across schools, with the proportion of pupils eligible for Free School Meals (FSM) ranging from 6% to 73.8%, suggesting participation from both relatively advantaged and highly disadvantaged contexts. Across the sample, the percentage of FSM pupils participating in the study was 34% (Table 2, Appendix G), compared to the national average of 25.7% (DfE, 2025). The sample also included a mix of school types, including community schools, academy converters, academy sponsor-led schools, and voluntary aided schools.

Table 3 also illustrates how recruitment patterns differed between the two recruitment phases. Phase 1 (April–July 2025) generated 43 EoIs, resulting in 15 signed MoUs (35%). These schools subsequently contributed 19 teachers who created accounts on *WhatWorked Teachers*, 15 of whom completed the full trial (79% completion). Phase 2 (September–October 2025) generated a further 15 EoIs, resulting in four signed MoUs (27%). These schools contributed six teachers who created accounts on *WhatWorked Teachers*, of whom two completed the full trial cycle (33% completion). Most attrition during Phase 2 occurred prior to trial implementation, with teachers either not progressing beyond randomisation or withdrawing before pre-test.

Participation also varied considerably between schools. Some schools contributed a single teacher, while others contributed multiple teachers across different year groups and trial designs (Table 4). Three schools had no teachers complete a full trial, although one of these schools contributed a teacher who reached randomisation. In contrast, several schools supported multiple completed trials. This variation illustrates why school-level MoUs could not be used to predict final teacher or pupil recruitment and why teacher-level platform engagement is a more informative feasibility measure.

Table 3: School characteristics

School	Location	Size	FSM	School Type	Teacher	Recruitment	Trial	Sample	Status
School 1	Urban	169	40.6%	Voluntary aided school	Teacher 1	Phase 1	Front-loaded LSCWC Year 5 Class	N/A	Incomplete: Randomisation
School 2	Urban	210	20.5%	Academy converter	Teacher 2	Phase 2	Front-loaded LSCWC Year 3 Class	N/A	Incomplete: Randomisation
					Teacher 3	Phase 2	Front-loaded LSCWC Year 4 Class	30	Incomplete: Pre-test
					Teacher 4	Phase 2	Front-loaded LSCWC Year 2 Small Group	20	Withdrawn: Pre-test
School 3	Rural	125	8%	Community school	Teacher 5	Phase 2	Front-loaded LSCWC Year 3 Class	25	Incomplete: Pre-test
School 4	Urban	441	73.8%	Academy converter	Teacher 6	Phase 1	Front-loaded LSCWC Year 5 Class	58	Withdrawn: Pre-test
					Teacher 7	Phase 1	Front-loaded LSCWC Year 5 Class	57	Withdrawn: Pre-test
					Teacher 8	Phase 1	Incremental LSCWC Variation Year 5 Class	58	Complete
School 5	Urban	266	15%	Community school	Teacher 9	Phase 1	Incremental LSCWC Variation Year 2 Small Group	5	Complete
					Teacher 10	Phase 1	Front-loaded LSCWC Year 3 Small Group	6	Complete
School 6	Urban	219	14.6%	Voluntary aided school	Teacher 11	Phase 1	Front-loaded LSCWC Year 4 Class	49	Complete
School 7	Urban	350	47.3%	Academy sponsor led	Teacher 12	Phase 1	Front-loaded LSCWC Year 2 Small Group	10	Complete
					Teacher 13	Phase 1	Front-loaded LSCWC Year 3 Class	45	Complete
					Teacher 14	Phase 1	Front-loaded LSCWC Year 4 Class	45	Complete
					Teacher 15	Phase 1	Front-loaded LSCWC Year 5 Class	41	Complete
School 8	Urban	374	15.6%	Community school	Teacher 16	Phase 1	Front-loaded LSCWC Year 5 Class	39	Complete
					Teacher 17	Phase 1	Front-loaded LSCWC Year 4 Class	39	Complete
School 9	Rural	203	14.3%	Community school	Teacher 18	Phase 1	Front-loaded LSCWC Year 2 Small Group	10	Complete
School 10	Urban	451	41.1%	Community school	Teacher 19	Phase 1	Front-loaded LSCWC Year 5 Class	16	Complete
School 11	Urban	430	6%	Community school	Teacher 20	Phase 1	Front-loaded LSCWC Year 5 Small Group	16	Complete
School 12	Urban	202	14.4%	Academy converter	Teacher 21	Phase 1	Front-loaded LSCWC Year 4 Small Group	28	Complete
School 13	Urban	237	20.3%	Voluntary aided school	Teacher 22	Phase 1	Front-loaded LSCWC Year 3 Small Group	8	Complete
School 14	Urban	178	41.2%	Community school	Teacher 23	Phase 1	Front-loaded LSCWC Year 4 Small Group	19	Complete

School 15	Rural	368	19.4%	Community school	Teacher 24	Phase 2	Incremental LSCWC Variation Year 5 Class	49	Complete
					Teacher 25	Phase 2	Front-loaded LSCWC Year 3 Class	49	Complete

Table 4: Distribution of schools and pupils in front-loaded and incremental LSCWC trials by design (class/small group)

Approach	**Design**	**Trials *N***	**School *N***	**Pupil *N***
Front-loaded	Class-Level Randomisation	8	5	323
	Small-Group Randomisation	7	7	97
Incremental	Class-Level Randomisation	2	2	107
	Small-Group Randomisation	1	1	5

Data collection

The study employed multiple methods to assess feasibility, including implementation evaluation surveys, interviews, and platform usage data. Key indicators and questions included:

Engagement and retention

- Rates of enrolment and teacher withdrawal prior to randomisation (platform usage data).
- Levels of retention and factors contributing to loss-to-follow-up (platform usage data and follow-up implementation survey, focus group/interview).

Teacher understanding and compliance

- Are teachers able to grasp the essential requirements of participating in an RCT, including randomisation, research consent, and data protection procedures? (implementation evaluation surveys, focus group/interview)
- To what extent do teachers comply with class or pupil allocations based on randomisation? (implementation evaluation surveys, focus group/interview)
- How does the remote trial management process influence teacher engagement and adherence to the trial protocol? (implementation evaluation surveys, focus group/interview)

Data collection and usability

- Is it feasible for teachers to administer outcome measures independently? (platform usage data, implementation evaluation surveys, focus group/interview)
- What additional guidance or resources do teachers require to ensure accurate data collection? (implementation evaluation surveys, focus group/interview)
- How effectively do teachers interpret and respond to automated impact reports generated by the platform? (implementation evaluation surveys, focus group/interview)

Data analysis

Platform usage data

Platform usage data, including enrolment rates, withdrawal rates, and retention levels, were summarised using descriptive statistics to identify trends and patterns. Compliance rates were calculated as the percentage of teachers who successfully completed pre- and post-test data entry within the study timeframe, alongside reviewing IPE survey data and interview responses to understand the extent to which teachers followed procedure.

Survey responses

All teachers who completed either the front-loaded or incremental LSCWC evaluations were invited to complete an online Implementation Process Evaluation (IPE) survey. However, although teachers completing the incremental evaluation were invited to participate, no responses were received from this group. Consequently, the implementation findings presented in this report relate exclusively to teachers who completed the front-loaded evaluation. Survey responses were analysed using descriptive statistical and qualitative approaches appropriate to the type of data collected. For closed-response items, frequency counts and percentages were calculated for each response option. Denominators were based on the number of non-missing responses for each item. For numeric items, summary statistics were produced, including the mean, standard deviation, median, and minimum and maximum values. Open-text responses were analysed using a descriptive content analysis approach. Responses were extracted verbatim, read in full by the lead researcher, and grouped according to recurring topics relevant to each survey question (for example, usability, barriers to implementation, and suggestions for improvement). Given that responses were generally brief, the analysis aimed to provide a descriptive summary of participants' comments rather than develop an interpretive thematic framework. Representative quotations were selected to illustrate commonly reported views where appropriate. Missing responses were identified and reported for each survey item.

Interview transcripts

Participating teachers were offered the option of either an individual interview or a focus group. All teachers who responded elected to participate in a one to one interview, indicating that this format was more practical given the demands of classroom timetables. All interview participants had completed the front-loaded LSCWC evaluation; although teachers completing the incremental evaluation were also invited to participate, no responses were received from this group. All interview participants were from different schools. Interview transcripts were analysed thematically. The lead researcher read each transcript several times to become familiar with the data before generating initial codes manually. Related codes were grouped into broader themes describing common barriers and facilitators to implementation. Themes were reviewed against the original transcripts to ensure they accurately reflected participants' responses. Given the relatively small number of interviews and the exploratory nature of this feasibility study, qualitative analysis software was not used. Particular attention was given to participants' understanding of randomisation procedures, ethical considerations, consent processes. and interpretation of the automated impact reports, reflecting the study's feasibility objectives.

Platform impact data

The *WhatWorked Teachers* platform automatically analyses pupil attainment data following completion of each classroom micro-RCT to generate an individual teacher impact report. Teachers upload anonymised pupil-level pre-test and post-test spelling scores (maximum score = 20). The platform then automatically runs an analysis of covariance (ANCOVA) using an R package to estimate the adjusted difference between intervention and control groups while controlling for baseline attainment. For each individual classroom trial, the platform fits the following model:

$$\text{Post-test} = \beta_0 + \beta_1(\text{Pre-test}) + \beta_2(\text{Treatment}) + \varepsilon$$

where treatment allocation represents the intervention effect after adjustment for pre-test attainment. The adjusted mean difference is used to calculate Hedges' g with a small-sample correction and corresponding 95% confidence intervals. Percentage gain is also calculated from the adjusted mean difference as a percentage of the adjusted control group mean and presented as the headline result within the teacher impact report because this metric is considered more accessible

for practitioners than standardised effect sizes. Complete mathematical specifications for the ANCOVA model, effect size calculations, and confidence intervals are provided in Appendix G.

The platform analyses each teacher's classroom trial independently and does not combine pupil data across schools when generating teacher reports. In addition to overall impact estimates, teachers receive contextual information including descriptive statistics for intervention and control groups, adjusted post-test means, percentage gain, effect sizes with confidence intervals, and graphical summaries of pupil progress. These outputs are intended to support teachers' interpretation of findings within the context of their own classroom rather than provide definitive estimates of intervention effectiveness.

For the purposes of the feasibility study, anonymised pupil-level data from completed classroom trials were exported from the platform and analysed collectively to evaluate the performance of the platform's analytical functionality rather than the efficacy of the spelling interventions themselves. Analyses followed the intention-to-treat (ITT) principle, whereby pupils were analysed according to their original randomised allocation irrespective of implementation fidelity. Primary analyses used linear mixed-effects models with post-test score as the outcome variable, treatment allocation and pre-test score entered as fixed effects, and school included as a random intercept to account for clustering of pupils within schools. Intraclass correlation coefficients (ICCs) were calculated to estimate the proportion of variance attributable to differences between schools. Baseline imbalance was assessed descriptively using pre-test attainment and FSM eligibility. Subgroup analyses explored whether estimated treatment effects differed according to FSM eligibility, trial format (small-group versus class-level randomisation), and year group using interaction terms within the mixed-effects models. As individual classroom micro-RCTs were not powered to detect subgroup effects, these analyses were considered exploratory.

Finally, the fifteen completed front-loaded classroom trial replications were synthesised using a cumulative random-effects meta-analysis. Unlike the teacher impact reports, which analyse each classroom independently, this analysis combined effect size estimates across replications to explore whether multiple classroom micro-RCTs could generate a stable pooled estimate while accounting for between-trial heterogeneity. Because proximal outcome measures were used and individual classroom samples were necessarily small, cumulative meta-analysis was considered a more appropriate indicator of intervention performance than individual classroom effect sizes alone. Missing outcome data was minimal (<5%) and was handled using complete-case analysis. Attrition rates were reported for both intervention and control groups. All statistical analyses were undertaken in R.

Timeline

Table 5: Timeline

Date	Activity
3rd March – 11th April 2025	Study plan IPE design
28th April – 18th July 2025	School recruitment for September start (Phase 1) Micro-RCTs created on the *WhatWorked Teachers* platform
2nd September – 24th October 2025	School recruitment for October start (Phase 2) Phase 1: randomisation, pre-test, six-week intervention delivery, post-test, IPE survey, focus groups/interviews
1st November – 18th December 2025	Phase 2: randomisation, pre-test, six-week intervention delivery, post-test, IPE survey, focus groups/interviews
18th December – 1st March 2026	Mop-up focus groups/interviews Draft feasibility report Secondary analysis Cumulative meta-analysis Final draft report

Findings

To what extent can teacher-led micro-RCTs serve as a practical, robust method for evaluating educational practices?

Fifteen primary schools participated in the feasibility study, exceeding the minimum recruitment target of twelve schools. Across completed trials, the mean time between teachers uploading pre-test and post-test scores was 47 days, indicating that evaluations were completed within the planned six-week implementation period (Appendix H, Table 1).

Pupil-level data demonstrated that the platform successfully supported completion of the evaluation cycle with high levels of outcome data capture. In the front-loaded LSCWC trials, 416 of 420 pupils (99.0%) completed the pre-test, and 413 of 416 pupils (99.3%) subsequently completed the post-test, representing an overall retention rate of 98.3% from randomisation to post-test. Retention was lower in the incremental LSCWC trials, where 106 of 112 pupils (94.6%) completed the pre-test, and 93 of 106 pupils (87.7%) completed the post-test, giving an overall retention rate of 83.0%. However, only three incremental trials were completed during the feasibility study and findings relating to this pathway should therefore be interpreted cautiously (Table 6). Overall, these findings suggest that teachers were able to implement classroom micro-RCTs and collect outcome data successfully within routine school practice.

Table 6: Attrition of pupils in front-loaded and incremental LSCWC trials

Stage	Front-loaded LSCWC Trials			Incremental LSCWC Trials		
	Total	Loss	Participated	Total	Loss	Participated
Randomised, no pre-test	420	4	416	112	6	106
Pre-test, no post-test	416	3	413	106	13	93

Survey findings were similarly positive regarding the practical delivery of the evaluation process. Across pooled class-level and small-group surveys[1], 8 of 11 respondents (72.7%) agreed that choosing a micro-trial was straightforward, 10 of 11 (90.9%) agreed that the platform was teacher-friendly, 10 of 11 (90.9%) agreed that scoring and data upload was manageable within workload, and 10 of 11 (90.9%) agreed that a six-week intervention timeline was realistic. Interview findings[2] were consistent with these responses. All six interview participants described the platform as straightforward to navigate, with a logical sequence of activities that enabled them to administer assessments, upload data, and complete each stage of the evaluation without substantial disruption to routine teaching. Teachers reported that the platform guided them clearly through the evaluation process and that the overall workflow was manageable alongside their existing classroom responsibilities. The same evidence shows that report interpretation was noticeably weaker than platform use. Only 4 of the 11 pooled survey respondents agreed that the automated report was easy to interpret, and only 6 of 11 agreed that the results felt relevant to their classroom context. Interviewees praised the speed and credibility of automated reporting, but one teacher described not knowing what they were looking at when the report appeared and asked for a short explanatory summary; another described different staff reading the same report differently. This is a highly relevant feasibility finding, as the platform can generate outputs, but not all users can interpret them confidently. Teacher-led micro-RCTs appear feasible as a practical method, but evidence is more promising for workflow completion than for interpretive confidence. The platform successfully operationalised the trial process; the weaker link was not basic usability but the accessibility of the analytic output.

Is it feasible to engage schools and teachers and get them to participate? What are the facilitators and barriers?
Overall, the findings indicate that engaging schools and teachers in teacher-led classroom micro-RCTs is feasible, although recruitment and participation were influenced by the timing of recruitment and competing school priorities. Recruitment

[1] For IPE survey data and respondent characteristics. Please see Appendix I.
[2] For IPE interview prompts and participant characteristics. Please see Appendix J.

had met the minimum recruitment target of 12 schools. Fifty-eight schools initially expressed interest in participating, 19 schools subsequently signed Memoranda of Understanding (MoUs), and 15 schools ultimately enrolled at least one teacher into a trial. Four schools that signed MoUs did not progress to teacher participation and did not respond to follow-up communication. Teacher recruitment followed a similar pattern. Twenty-five teachers enrolled onto the platform, 23 commenced a trial and 18 completed the full evaluation cycle. Most attrition therefore occurred before trial implementation began rather than during delivery of the evaluation itself. Once teachers reached the implementation stage, retention was exceptionally high, with all teachers who completed pre-testing subsequently completing the intervention and post-test data collection. This suggests that the principal engagement challenge was converting initial institutional commitment into active teacher participation rather than sustaining teacher engagement once implementation had begun.

Recruitment took place in two phases. During the first phase, 43 schools expressed interest, resulting in 15 signed MoUs (34.9%) and 19 participating teachers, of whom 15 completed a trial (78.9%). A second recruitment phase was undertaken, as it was unclear how many teachers from each participating trial would run an evaluation. To ensure recruitment targets could be met, a second phase of recruitment was initiated. Among teachers who participated, the principal facilitators of engagement were the perceived relevance of the platform to classroom practice, the manageable workload associated with implementation, and the opportunity to generate evidence using their own pupils. Survey and interview findings consistently indicated that teachers viewed the evaluation process as compatible with routine classroom practice. One teacher described the trial as '*so simple but so useful… it just fitted into what we were already doing*,' illustrating the importance of embedding evaluation within existing teaching routines. Teachers also valued receiving rapid feedback and reported sharing findings with colleagues and senior leaders, suggesting that engagement extended beyond individual participants and stimulated wider professional discussion about evidence-informed practice.

Barriers to engagement were predominantly practical rather than conceptual. Schools cited timetable pressures, staffing changes, and class reorganisation as the primary reasons for withdrawing before implementation. Teachers also reported some uncertainty during the initial trial selection process when choosing between available evaluation designs, suggesting that additional onboarding guidance may improve confidence during recruitment. Administrative burden created some friction where teachers perceived duplication between spreadsheet-based processes and platform data entry, while several participants suggested that providing ready-made teaching resources, such as spelling materials, would further reduce preparation time. These findings suggest that the principal challenge for future implementation is improving the conversion of school-level interest into active teacher participation rather than maintaining engagement once teachers have begun using the platform. Teacher retention during implementation was exceptionally strong, indicating that once teachers started a classroom micro-RCT, they were highly likely to complete it. Future scale-up efforts should therefore focus on optimising recruitment timing, simplifying onboarding, and reducing administrative barriers during the period between school commitment and trial commencement.

Overall, engagement is feasible when the trial is perceived as professionally useful and manageable, but remains sensitive to workload, clarity of onboarding, and local organisational conditions. These findings should be interpreted with some caution because participation in the implementation surveys and interviews was voluntary and may therefore be subject to self-selection bias. In addition, the feasibility study did not examine continued use of the platform beyond the trial period, so it was not possible to assess longer-term teacher engagement or sustained platform adoption.

Is it feasible to get teachers to select a trial of interest, comply/adhere to protocols and implement the intervention? What are the facilitators and barriers?

Teachers were able to select an evaluation, implement the intervention, and adhere to the trial procedures, although implementation challenges differed between class-level and small-group delivery models. Adherence to the allocated trial design was high. All four teachers delivering class-level evaluations reported implementing the intervention as originally allocated, whereas four of seven teachers delivering small-group evaluations reported making minor adaptations after randomisation. These adaptations were principally made in response to practical classroom considerations, including SEND (special educational needs and disabilities) provision, pupil absence, timetabling constraints, and behaviour management, rather than dissatisfaction with the trial procedures themselves.

Teachers generally reported that they were able to maintain separation between intervention and comparison groups; however, perceived contamination was greater within small-group trials than class-level trials. Three out of four class-level

teachers agreed or strongly agreed that they had avoided introducing intervention elements to the comparison group, compared with two of seven teachers delivering small-group evaluations. Open-text responses suggested that these ratings reflected the practical realities of teaching spelling rather than deliberate contamination of trial allocation. Teachers described unavoidable overlap between intervention and comparison groups through shared spelling rules, common word lists, and whole-class teaching activities that formed part of routine classroom practice. Descriptions of business-as-usual spelling instruction supported this interpretation. Comparison pupils commonly received explicit spelling teaching involving repeated word practice, spelling-rule instruction, homework activities, and in one class-level trial, use of the 'Look, Say, Cover, Write, Check' approach. Teachers' perceptions of similarity between usual practice and the intervention nevertheless varied. Class-level teachers generally considered their existing practice to be at least 'a little similar' to the intervention, whereas small-group teachers more frequently viewed the intervention as distinct from their routine teaching. This suggests that teachers differentiated the intervention primarily through its structure, organisation, and emphasis rather than the complete absence of shared instructional components.

A key facilitator of implementation was teachers' understanding of randomisation and its purpose. Interview data shows that teachers recognised random allocation as essential for a fair test and generally adhered to it, even where it conflicted with professional instinct. As one teacher explained, '*if you changed the groups… it defeats the purpose*,' reflecting strong conceptual buy-in to the trial design. Implementation was also supported by clear platform guidance, structured workflows, and the ability to integrate the intervention into existing teaching routines. Most teachers reported delivering sessions broadly in line with the protocol (75% class-level; 85.7% small-group), indicating that the intervention itself was feasible to enact within classroom conditions. However, small-group trials introduced additional logistical complexity. Staffing or cover was reported as a barrier by 5/7 small-group teachers, compared to none in class-level trials, and timetabling clashes were also more common. These challenges reflect the difficulty of maintaining stable group allocation and separation within a single classroom context.

These findings suggest that implementing teacher-led classroom micro-RCTs is feasible in routine school settings, but that maintaining complete separation between intervention and comparison conditions may be more challenging where both groups are taught within the same classroom. Future iterations of the platform should therefore provide clearer guidance and practical examples to help teachers minimise potential contamination in within-class evaluations, while recognising that some overlap between intervention and usual classroom practice is unavoidable for curriculum-based educational interventions.

Is it feasible to draw conclusions from the data gathered (including completeness, reliability, effect sizes etc.)? What are the challenges and can they be mitigated?

The feasibility study demonstrated that teachers could successfully generate analysable pupil-level data through the *WhatWorked Teachers* platform. Data completeness was high among teachers who commenced implementation. In the front-loaded evaluations, 413 of 420 randomised pupils (98.3%) completed the post-test assessment, while 93 of 112 pupils (83.0%) completed the post-test in the incremental evaluations. At teacher level, all 18 teachers who entered pre-test data subsequently uploaded post-test results, indicating that complete outcome datasets were routinely obtained once implementation had begun.

The platform successfully transformed this data into automated analytical outputs. For each completed classroom trial, teachers received an impact report containing adjusted comparisons between intervention and control groups, percentage gain, standardised effect sizes, and associated confidence intervals. Across the pooled front-loaded evaluations, the estimated effect size was 0.27 (95% CI: 0.03 to 0.51), based on 15 completed classroom trials and 413 pupils. The pooled incremental evaluations produced an estimated effect size of 0.70 (95% CI: 0.37 to 1.03), based on three completed trials and 93 pupils. Detailed analyses for individual classroom trials, year-group aggregations, and FSM subgroup analyses are presented in Appendix G.

However, the ability to generate statistical outputs should not be equated with the ability to draw robust conclusions from them. Individual teacher-level evaluations were based on relatively small numbers of pupils and consequently produced effect estimates that were often imprecise, with wide confidence intervals. The achieved sample was also unevenly

distributed across trial variants, year groups, and delivery models, particularly for the incremental evaluations, limiting the precision of subgroup analyses and reducing confidence in comparisons between conditions. Although an a priori sample size calculation informed the planned recruitment assumptions, the feasibility study was not powered to detect effects within individual classroom trials or exploratory subgroup analyses. These findings should therefore be interpreted as exploratory rather than definitive.

Reliability was considered in terms of both data completeness and the consistency of findings across trial replications. Outcome data capture was consistently high once implementation commenced, demonstrating that teachers were able to collect and submit complete datasets using the platform. Nevertheless, variation in estimated effects across individual classroom trials, together with intraclass correlation coefficients ranging from 0.15 to 0.38, indicated that school context and implementation differences contributed meaningfully to variation in outcomes. This variability reinforces the importance of interpreting evidence across multiple trial replications rather than relying on findings from a single classroom evaluation. The reliability of the spelling outcome measure itself was not assessed as part of this feasibility study.

The study also identified an important challenge regarding interpretation of the analytical outputs. While teachers consistently reported that entering pupil data and generating reports was straightforward, confidence in interpreting those reports was substantially lower. Only four of eleven survey respondents agreed that the automated impact report was easy to interpret, while interview participants described uncertainty about the meaning of effect sizes, confidence intervals, and the practical implications of the findings. Several participants requested a Plain English summary alongside the statistical outputs, and one participant noted that different members of staff interpreted the same report differently. These findings suggest that the platform currently succeeds in automating statistical analysis but is less successful in supporting teachers to interpret the resulting evidence confidently and consistently.

The findings therefore highlight an important distinction between analytical feasibility and decision-making feasibility. The platform was able to generate complete datasets and appropriate statistical analyses from teacher-entered data, demonstrating that automated classroom evaluation is technically feasible. However, the evidence generated by individual classroom trials was often insufficiently precise to support strong conclusions, and teachers were not always confident in interpreting the outputs or determining what actions should follow. This represents a key design challenge for future platform development. Several strategies could mitigate these limitations. Statistical precision could be improved by increasing the number of classroom trial replications, achieving a more balanced distribution across trial variants and year groups, and placing greater emphasis on cumulative evidence rather than individual classroom estimates. Equally important, the presentation of results should be refined to improve accessibility for teachers. This may include providing Plain English summaries of findings, clearer explanations of effect sizes and confidence intervals, explicit guidance regarding the uncertainty associated with small classroom trials, and recommendations about when findings should inform classroom practice and when further evidence should be sought.

The findings demonstrate that it is feasible to generate, analyse, and synthesise classroom evaluation data using the *WhatWorked Teachers* platform. However, the feasibility of drawing reliable and actionable conclusions is more limited. The principal constraints are not data completeness or the platform's analytical capability, but the statistical uncertainty inherent in small classroom trials and teachers' confidence in interpreting the resulting evidence. Future development should therefore focus not only on expanding the cumulative evidence base but also on improving how statistical findings are communicated so that teachers can make appropriate evidence-informed decisions.

What are the typical effect sizes to be expected when micro-RCTs are utilised in an educational context? Does the use of micro-RCTs have the potential to produce cumulative evidence through meta-analysis?

The feasibility study demonstrated that the *WhatWorked Teachers* platform can successfully generate standardised effect size estimates from teacher-led classroom micro-RCTs. Across the pooled front-loaded evaluations, the estimated effect size was 0.27 (95% CI: 0.03 to 0.51), while the pooled incremental evaluations produced an estimated effect size of 0.70 (95% CI: 0.37 to 1.03). However, these estimates were derived from markedly different sample sizes (413 versus 93 pupils),

and the incremental estimate was based on only three completed trials. Consequently, the findings do not provide sufficient evidence to establish a typical effect size for classroom micro-RCTs and should be interpreted as exploratory.

The findings also illustrate an important characteristic of classroom micro-RCTs. Individual teacher-led evaluations are typically conducted with relatively small numbers of pupils and are therefore unlikely to produce highly precise estimates unless intervention effects are comparatively large. This was reflected in the wide confidence intervals observed for many individual classroom trials (Appendix K), indicating considerable statistical uncertainty around some estimates. These findings do not diminish the value of individual classroom evaluations but suggest that their primary purpose is to support teacher enquiry and professional reflection within a specific classroom context rather than to provide definitive evidence of intervention effectiveness. This distinction was reflected in teachers' experiences of the automated impact reports. Teachers consistently valued receiving immediate feedback following completion of their evaluations and viewed the reports as professionally useful. As one participant explained, *'I could see that the intervention had made enough of an impact to share it with SLT,'* illustrating how the reports stimulated professional discussion and supported evidence-informed reflection within schools. However, teachers were less confident interpreting statistical outputs such as effect sizes and confidence intervals, suggesting that future platform development should provide clearer guidance regarding the strength and uncertainty of individual trial findings.

The study also demonstrates that the platform provides the technical infrastructure required to generate cumulative evidence across multiple classroom evaluations. By standardising outcome measures, automating analyses, and calculating effect sizes using consistent analytical methods, findings from independent classroom micro-RCTs can be synthesised. A random-effects meta-analysis of the fifteen completed front-loaded evaluations produced a pooled effect estimate of 0.35 (95% CI: 0.03 to 0.67). Although individual classroom effect sizes ranged from -1.12 to 1.30, statistical heterogeneity was low (I^2 = 0%), suggesting that variation between studies was largely consistent with sampling variation rather than systematic differences in intervention effects. This demonstrates that meaningful cumulative synthesis is achievable when classroom evaluations are conducted using common protocols and analytical procedures.
The principal strength of the *WhatWorked Teachers* platform lies in its dual function. At the individual teacher level, the platform provides rapid, classroom-specific feedback that can support professional enquiry, reflection, and discussion about teaching approaches. At the system level, the platform enables comparable classroom evaluations to accumulate into a growing body of evidence that can be synthesised through meta-analysis to produce more precise and generalisable estimates than are possible from individual classroom trials alone. Future development should strengthen both functions. At the classroom level, automated reports should place greater emphasis on explaining statistical uncertainty in Plain English and provide clearer guidance on how teachers should interpret and act upon findings from a single trial. At the system level, continued accumulation of classroom evaluations across a wider range of schools, year groups, and intervention contexts will increase statistical precision and enable more robust cumulative evidence to inform educational practice. While individual classroom micro-RCTs are best viewed as supporting evidence-informed professional enquiry, the greatest long-term potential of the platform lies in systematically building a cumulative evidence base through repeated classroom evaluations.

Discussion

This feasibility study demonstrates that teacher-led classroom micro-randomised controlled trials can be implemented through the *WhatWorked Teachers* platform within routine primary school practice. Teachers were generally able to recruit pupils, administer assessments, upload data, and complete evaluations within the planned timeframe, and the platform successfully generated automated impact reports from teacher-entered data. High levels of pre-post-test completion were achieved among teachers who commenced implementation, mirroring those observed in the earlier North East Combined Authority pilot (WhatWorked Education, 2025), despite the current study involving schools recruited nationally and supported through a lighter-touch onboarding model. This consistency suggests that the operational processes underpinning the platform, including trial delivery, outcome data collection, and report generation, can be implemented successfully across different educational settings. However, the study also identified important constraints on the conclusions that can be drawn from individual classroom evaluations. Although the platform proved technically capable of supporting teacher-led experimentation, the findings highlight challenges relating to statistical precision, potential

contamination within some trial designs, and teachers' interpretation of analytical outputs. Consequently, the principal contribution of the study is not simply that classroom micro-RCTs can be conducted, but that it identifies the conditions under which they are most likely to generate useful and trustworthy evidence.

This distinction is particularly important within the current English education policy context, where schools are increasingly expected to make evidence-informed decisions to improve pupil attainment and reduce educational disadvantage (Leckie and Goldstein, 2019; Prior *et al.*, 2021; Parliamentary Office of Science and Technology, 2024). Teacher-led evaluation has the potential to provide schools with timely evidence that is directly relevant to their own pupils and contexts. However, the present findings suggest that the value of this evidence lies in supporting structured professional enquiry and local improvement rather than providing definitive estimates of intervention effectiveness from individual classroom trials. This reflects the statistical realities of small-scale educational evaluations, where limited sample sizes inevitably constrain precision.

The findings both support and extend the existing literature on teacher-led experimentation. Churches (2016) argued that schools possess the capacity to engage with rigorous evaluation methods and that teachers are not inherently resistant to concepts such as control groups, randomisation, or statistical reasoning. The present study broadly supports this conclusion. Teachers demonstrated a good understanding of random allocation, generally adhered to trial procedures, and regarded participation as professionally worthwhile. However, this study also extends the literature by showing that engagement with research methods does not necessarily translate into confidence in interpreting statistical findings. While teachers found the platform straightforward to use and valued the immediate availability of automated reports, many reported uncertainty when interpreting effect sizes, confidence intervals, and the practical implications of the findings. This suggests that future development should focus not only on simplifying the process of conducting classroom evaluations but also on improving how statistical evidence is communicated to support informed professional judgement.

The study also highlights an important methodological challenge for teacher-led classroom evaluations. Although outcome data was collected successfully and implementation fidelity was generally high, maintaining complete separation between intervention and comparison conditions was more challenging in within-class, small-group evaluations. Teachers described unavoidable overlap between intervention and usual classroom practice, particularly where common spelling rules, word lists, or whole-class teaching activities were shared across groups. While this does not invalidate the methodology, it demonstrates that some degree of pedagogical contamination is difficult to avoid in curriculum-based educational research and should be considered when designing future classroom trials. Greater implementation guidance and clearer examples of strategies for minimising contamination may improve internal validity, particularly for within-class designs.

Finally, the study suggests that the greatest long-term potential of the *WhatWorked Teachers* platform lies in its ability to combine many comparable classroom evaluations into a cumulative evidence base. Individual classroom trials provided teachers with rapid, context-specific feedback that supported professional reflection, but the study also demonstrated the statistical limitations of relying on small, isolated evaluations. Wide confidence intervals and variability in individual effect estimates limited the strength of conclusions that could be drawn from single trials. In contrast, the standardised analytical approach adopted by the platform enabled findings to be synthesised across multiple replications through meta-analysis, offering a pathway towards more precise and generalisable estimates as the evidence base grows. The platform should therefore be viewed not simply as a mechanism for generating isolated teacher evaluations, but as an infrastructure for accumulating high-quality classroom evidence over time.

Engagement and participation

The findings suggest that recruiting schools and teachers to participate in teacher-led classroom micro-randomised controlled trials is feasible, but sustaining engagement throughout the complete evaluation process remains a key challenge. Recruitment required two phases owing to the combined challenges of recruiting late in the summer term, uncertainty regarding teacher participation following school-level agreement, and the operational realities of recruiting schools during the opening weeks of a new academic year. Despite these challenges, the study exceeded its predefined school recruitment target (15 schools compared with a target of 12) and came close to achieving its teacher recruitment target (23 teachers randomised compared with a target of 24). However, considering engagement across the entire platform

workflow provides a more nuanced assessment of feasibility. Following the removal of duplicate accounts and senior leaders exploring the platform, 29 teachers created platform accounts. Of these, 23 progressed to randomisation, and 18 completed the full trial cycle, representing an overall completion rate of 62% from initial registration. Although this demonstrates that a substantial proportion of teachers were willing to engage with the platform, it also highlights attrition before and during implementation, suggesting that improving progression from initial registration to successful trial completion should be a priority for future development.

Teachers who discontinued participation before completing a trial were contacted by email and invited to provide feedback regarding their reasons for withdrawing; however, the response rate was low, limiting the conclusions that could be drawn. The limited feedback received, together with observations made during recruitment and implementation, suggested that competing priorities, staffing changes, timetable pressures, and uncertainty surrounding intervention selection contributed to disengagement. These findings are consistent with the wider implementation literature, which suggests that successful adoption depends not only on the perceived value of an innovation but also on its compatibility with schools' priorities, capacity, and day-to-day working practices. The Education Endowment Foundation's *A School's Guide to Implementation* (Sharples, Eaton, and Boughelaf, 2024) similarly argues that schools should identify a clear problem, assess the fit and feasibility of an approach, and establish supportive implementation conditions before expecting sustained engagement. In the present study, interventions were selected by the research team to evaluate the feasibility of the methodology rather than by participating teachers. While this ensured consistency across trials, it may have reduced motivation for teachers whose immediate classroom priorities differed from the interventions offered. Given the relatively short duration and low administrative burden of classroom micro-RCTs, allowing teachers or schools to evaluate interventions that address their own improvement priorities may increase engagement by strengthening the perceived relevance of participation.

Among teachers who completed the evaluation cycle, perceptions of the platform were consistently positive. Participants described the platform as straightforward to use, professionally relevant, and valuable because it generated timely findings that could be shared with colleagues and school leaders. These findings align with wider evidence suggesting that practitioner enquiry is most likely to succeed when it addresses authentic classroom problems, is supported through collaborative professional cultures, and fits naturally within existing school improvement processes (Cain, 2015; Leat, Reid, and Lofthouse, 2015; Owen, Watkins, and Hughes, 2022). Similarly, Preston, Woods, and Qualter (2026) identify leadership support, collaboration, and manageable implementation demands as key facilitators of engagement, with competing priorities and unnecessary complexity acting as common barriers. The present findings closely mirror this pattern, suggesting that teachers were motivated less by the novelty of the platform than by its potential to answer meaningful classroom questions using evidence from their own pupils.

The barriers identified during the feasibility study, including uncertainty when selecting trials, duplication between spreadsheet and platform data entry, and requests for ready-made implementation materials, are particularly important because they arise at the point where teachers decide whether continued participation is worthwhile. Several design refinements have already been implemented to address these issues (p. 45). Future research should evaluate whether these modifications improve progression through the platform and increase completion rates, while continuing to monitor the balance between reducing user burden and maintaining methodological rigour.

Implementation fidelity and design choice

A particularly encouraging finding was that teachers were not only willing to use randomisation but generally understood why it was important. Concerns surrounding teacher-led trials often assume that practitioners will override random allocation on pastoral or intuitive grounds, thereby undermining internal validity. In contrast, interview participants consistently described randomisation as the basis of a fair comparison and demonstrated a commitment to preserving that principle throughout the evaluation. These findings support earlier work suggesting that schools are not inherently resistant to control groups, experimental designs, or statistical reasoning (Churches, 2016). They also indicate that, under low-stakes conditions, the principles underpinning rigorous evaluation can be compatible with professional judgement rather than perceived as conflicting with it.

However, the study also demonstrates that successful randomisation alone does not guarantee implementation fidelity. Maintaining clear separation between intervention and comparison conditions proved considerably more challenging,

particularly within small-group designs where both groups remained within the same classroom environment. Class-level evaluations demonstrated stronger fidelity across almost every implementation indicator. No class-level respondent reported post-randomisation changes, compared with three of seven teachers conducting small-group trials. Similarly, perceived contamination was lower in class-level trials (three of four respondents agreed contamination had been successfully avoided compared with two of seven in small-group trials), while staffing and timetabling difficulties were reported exclusively by teachers implementing small-group designs. These findings suggest that although both approaches are operationally feasible, they present different implementation challenges.

It is important to recognise that complete separation between intervention and comparison conditions is rarely achieved, even in conventional educational randomised controlled trials. Educational interventions are implemented within complex school environments where teachers, pupils, and resources interact continuously, making some degree of contamination an acknowledged challenge rather than an exceptional occurrence. The findings of the present study therefore do not identify a methodological weakness unique to teacher-led classroom micro-RCTs. Instead, they suggest that the practical difficulties associated with maintaining distinct treatment conditions may be amplified within small-scale classroom evaluations, particularly where intervention and comparison pupils remain within the same classroom and where intervention activities closely resemble routine teaching practices.

The interview data supports this interpretation. Teachers did not describe intentionally deviating from the protocol or deliberately delivering intervention activities to comparison pupils. Rather, contamination appeared to arise naturally through everyday classroom practice, including shared spelling rules, common word lists, whole-class teaching, and routine spelling activities. Descriptions of business-as-usual instruction also revealed substantial overlap with the intervention itself. Many teachers reported routinely using repeated spelling rehearsal and *look, say, cover, write, check* activities as part of their normal teaching, meaning that the distinction between intervention and comparison conditions was sometimes less pronounced than intended. This highlights an important consideration when selecting interventions for classroom micro-RCTs: approaches that are clearly distinguishable from existing classroom practice are likely to produce stronger treatment contrast than interventions that share many of the same instructional components.

Although contamination is recognised as a threat to internal validity across educational research more broadly, its consequences may be proportionally greater in classroom micro-RCTs because individual evaluations are typically conducted with relatively small numbers of pupils. Even modest reductions in treatment contrast may reduce the likelihood of detecting genuine intervention effects, contributing to the wide confidence intervals observed in many of the individual classroom evaluations reported in this study. These findings reinforce the importance of interpreting implementation fidelity alongside statistical findings and suggest that intervention selection and trial design are critical determinants of the quality of evidence generated. Nevertheless, the greater implementation challenges associated with small-group designs should not necessarily preclude their use. Small-group randomisation enables participation by single-form entry schools that cannot randomise whole classes and allows teachers to evaluate interventions targeted at specific groups of pupils, such as lower-attaining learners or those requiring additional support. This is particularly relevant given the longstanding underrepresentation of small, rural, and disadvantaged schools in large-scale educational research (Hargreaves, 2009; Parker *et al*., 2021; McGlade and Kelly, 2025). Excluding small-group designs would therefore reduce opportunities for these schools to contribute to and benefit from evidence generation. Instead, the findings suggest that small-group evaluations require additional implementation safeguards and may be better suited to interventions that can be delivered with clearer separation from business-as-usual practice.

Future development of the platform should therefore focus on strengthening implementation fidelity rather than restricting the range of evaluation designs available to teachers. This may include providing clearer guidance on selecting interventions that are sufficiently distinct from existing classroom practice, offering practical strategies for minimising contamination, improving implementation planning tools, and supporting teachers to identify trial designs that best match both their educational question and classroom context. Future research should also investigate whether contamination varies across different subject areas and intervention types and identify the circumstances under which classroom micro-RCTs can achieve the greatest balance between methodological rigour and practical feasibility.

Outcome data and effect estimates

Although estimating intervention effects was not the primary purpose of this feasibility study, an important finding was that teachers were able to generate complete and analysable outcome datasets through the platform. Front-loaded LSCWC trials randomised 420 pupils and retained post-test data for 413 (98.3%), while incremental LSCWC trials randomised 112 pupils and retained post-test data for 93 (83.0%). At teacher level, all 18 teachers who entered pre-test data also entered post-test data. These findings demonstrate that classroom outcome data can be collected reliably through a teacher-led digital platform when data collection is embedded within routine classroom practice and supported by simple workflows. Together with the similarly high completion rates observed in the earlier North East Combined Authority pilot (WhatWorked Education, 2025), this provides encouraging evidence that the operational aspects of data collection are reproducible across different implementation contexts. The platform also successfully automated statistical analyses and produced immediate impact reports, substantially reducing the analytical burden placed on teachers. This addresses an important limitation identified in the wider literature on learning analytics, where teachers frequently report difficulties analysing educational data and translating findings into classroom decisions (Hase and Kuhl, 2024). Nevertheless, the present study suggests that automation shifts rather than removes the challenge. While teachers valued the speed, credibility, and accessibility of automated reports, confidence in interpreting statistical outputs was noticeably lower than confidence in interpreting simpler descriptive summaries such as percentage differences between groups. Several interview participants reported uncertainty regarding the meaning of effect sizes and confidence intervals, despite expressing confidence in the credibility of the reports themselves.

This finding has important implications for future platform development. Automated analyses may increase confidence in evidence, but unless the outputs are presented in ways that accurately communicate both findings and uncertainty, there is a risk that teachers could place greater confidence in conclusions than the data justifies. This raises an important ethical as well as methodological consideration. Automated impact reports should support evidence-informed professional judgement rather than encourage definitive conclusions. Future development should therefore prioritise user testing of alternative report formats, including Plain English narrative summaries, visual representations of uncertainty, explicit guidance on interpreting confidence intervals, and clear explanations of the circumstances under which findings should be regarded as exploratory rather than sufficient to inform substantive changes to classroom practice. Such refinements should be evaluated with teachers to ensure that reports are understandable without becoming misleading through oversimplification.

The feasibility study also provides important insights into the statistical limitations of teacher-led classroom micro-RCTs. The pooled front-loaded evaluations produced an estimated effect size of 0.272 (95% CI: 0.032 to 0.513), while the pooled incremental evaluations produced an estimated effect size of 0.699 (95% CI: 0.367 to 1.032). Although these represent the observed effect size estimates, the confidence intervals surrounding them remained wide, even after pooling data across multiple classroom evaluations. Individual classroom trials were characterised by substantially greater uncertainty, reflecting the relatively small numbers of pupils contributing to each evaluation. Consequently, the principal finding from the outcome analyses is not the magnitude of the point estimates themselves but the limited statistical precision with which those estimates could be determined. These findings illustrate an important distinction between the operational feasibility of classroom micro-RCTs and their statistical capability. The platform successfully generated valid statistical analyses from teacher-entered data, but the underlying statistical limitations associated with small-scale randomised evaluations remain unchanged. Unless intervention effects are comparatively large, individual classroom micro-RCTs are unlikely to generate sufficiently precise estimates to support high-level conclusions about intervention effectiveness. Consequently, the findings suggest that individual classroom impact reports are best regarded as supporting structured professional enquiry rather than providing definitive evidence for high-stakes instructional decision making. This interpretation is consistent with methodological guidance on pilot and feasibility studies, which cautions against drawing strong conclusions from treatment-effect estimates generated from small preliminary samples because such estimates are inherently unstable and potentially misleading (Sim, 2019).

The observed effect size estimates are broadly consistent with the wider literature demonstrating benefits of explicit spelling instruction and retrieval-based learning strategies (Graham and Santangelo, 2014; Agarwal, Nunes, and Blunt, 2021; Mawson and Kang, 2025; Dymock and Nicholson, 2017; Fisher, Cozens, and Greive, 2007). However, this consistency

should not be interpreted as confirmation of intervention effectiveness within the present study. Rather, it indicates that the observed estimates are plausible within the existing evidence base. The feasibility study was not designed or powered to determine whether the LSCWC approach is effective, whether one variant is superior to the other, or whether subgroup differences represent genuine variation rather than sampling error.

Taken together, these findings suggest that the value of the *WhatWorked Teachers* platform lies in enabling teachers to undertake structured, evidence-informed enquiry using data from their own classrooms. Secondly, by applying consistent analytical methods across many classroom evaluations, it provides the infrastructure for building a cumulative evidence base through repeated replication and synthesis. Future research should therefore focus not only on increasing the number of classroom evaluations contributing to cumulative analyses but also on optimising how statistical findings are communicated so that teachers understand both the insights and the limitations of the evidence generated.

Limitations

This study was a proof-of-concept feasibility evaluation of the *WhatWorked Teachers* platform. It was not an evaluation of the effectiveness of the spelling interventions teachers could test via the platform. The primary aim was to determine whether teachers could successfully use the platform to recruit pupils, conduct classroom randomisation, implement intervention and comparison conditions, collect outcome data, receive automated analyses, and engage with the resulting evidence within routine classroom practice. Consequently, the study was designed to establish the operational and methodological feasibility of the platform rather than to determine intervention efficacy. Although exploratory effect size estimates were generated successfully, these were included to demonstrate that the complete analytical workflow, from randomisation through to automated reporting, could operate within authentic school settings. The study was neither designed nor intended to establish whether the interventions were effective, estimate a typical intervention effect, or compare the relative effectiveness of the front-loaded and incremental approaches. The statistical findings should therefore be interpreted as evidence that the platform can successfully generate and report outcome analyses, rather than as evidence for or against the interventions themselves.

Implementation evidence was incomplete. Eleven teachers completed the implementation process evaluation survey (61% response rate), representing a range of year groups, school types, school sizes, and levels of disadvantage (Appendix I). However, all respondents participated in the front-loaded LSCWC condition during Phase 1 of delivery. Similarly, although interview participants represented a range of educational contexts, all had delivered the front-loaded intervention, with no representation from teachers completing the incremental variation. Urban schools were also overrepresented within both the survey and interview samples, while academy schools were not represented within the interview sample. Consequently, implementation findings relating to the incremental design remain tentative, and meaningful comparisons between delivery models cannot be made without further implementation evidence.
The study also provides only limited insight into why some teachers did not progress through the complete platform workflow. While teacher engagement was monitored throughout recruitment and implementation, relatively little information was obtained from teachers who withdrew before completing a trial. Follow-up emails invited participants to provide feedback regarding their reasons for discontinuation, but response rates were low. As a result, it was not possible to establish whether attrition reflected competing school priorities, intervention selection, workload, platform usability, or other contextual factors.

Implementation fidelity was assessed primarily through teacher surveys and interviews rather than independent classroom observation. This approach was appropriate for a feasibility study concerned with understanding teachers' experiences of using the platform and identifying practical implementation challenges. However, reports of treatment separation, contamination, and implementation adaptations represent teachers' accounts of their classroom practice rather than independently verified observations. Future evaluations could strengthen implementation evidence by incorporating classroom observations or other objective measures of fidelity where this is practical and proportionate.

An important finding emerging from the feasibility study concerns the interpretation of automated analytical reports. Although teachers consistently valued receiving rapid statistical feedback and regarded the reports as credible, confidence in interpreting effect sizes, confidence intervals, and statistical uncertainty was noticeably lower than confidence in

interpreting simpler descriptive summaries. This suggests that further development of the reporting interface is required before wider implementation. Future work should evaluate alternative approaches to communicating statistical findings, including Plain English summaries, visual presentation of uncertainty, and clearer guidance regarding the circumstances under which findings should be regarded as exploratory rather than definitive.

Finally, the study was designed, implemented, and evaluated by the organisation that developed the *WhatWorked Teachers* platform. While this facilitated rapid iteration of platform features throughout the feasibility study, it also introduces the possibility of investigator bias in study design, interpretation, and reporting.

Taken together, these limitations do not detract from the central purpose of the study, which was to establish whether teacher-led classroom micro-randomised controlled trials can be delivered successfully through a digital platform under everyday school conditions. Instead, they identify the key areas requiring further refinement before wider implementation, including strengthening teacher engagement throughout the platform workflow, improving implementation evidence across different trial designs, enhancing the communication of automated statistical findings, and undertaking independent evaluation. As such, the study has fulfilled its proof-of-concept purpose by demonstrating both the practical feasibility of the platform and the principal methodological considerations that should inform its future development and evaluation.

Contribution to the evidence base

Compared with the much more developed health literature on micro-randomised trials (Klasnja *et al*., 2015; Qian *et al*., 2022; Liu *et al*., 2017), educational applications remain relatively sparse (Breitwieser *et al.*, 2024). Persistent gaps between educational research and classroom practice mean that teachers often experience difficulties translating externally produced evidence into contextually appropriate action (Rickinson *et al*., 2019; OECD, 2023; Levin, 2013). Work on evidence-informed schools has repeatedly argued that research use depends on teacher capacity, school culture, and practical organisational structures (Brown and Zhang, 2016; Brown and Greany, 2018; Nelson and Campbell, 2017; Cain, 2015). This study contributes to the evidence base by demonstrating that automated, platform-based micro-trials may help operationalise those conditions: they reduce some of the administrative friction of formal research, preserve the core logic of randomised comparison, and produce outputs quickly enough to be used in school decision making. In this manner, the platform is not simply a data-collection tool, but an infrastructure for practitioner-led evidence creation and mobilisation.

Research on teachers' engagement with evidence suggests that research is most useful when it is translated into contextualised professional knowledge through collaborative interpretation and professional judgement, rather than treated as a set of universally transferable rules (Levin, 2011; Levin, 2013; Cain, 2015). The interview evidence presented in this study reinforces this message. Teachers used the impact reports generated by the platform to guide continuous professional development (CPD), discuss practice with senior leaders, and plan wider adoption. The study therefore contributes an important finding to the evidence-informed practice literature by showing that teachers can not only consume evidence, but that they can produce bounded, locally credible evidence when methodological and technical burden is reduced. This positions the platform-based model closer to a genuinely school-led evidence ecology rather than a one-way transfer model from university or external evaluator to school.

The study also provides a credible answer to the question of replication and cumulative evidence generation. One longstanding criticism of educational research is the lack of direct replication studies within the field, which limits the development of cumulative and robust evidence for practice and policy (Makel and Plucker, 2014; Perry *et al*., 2022; Cantley, 2023). Micro-randomised trial methodology emphasises standardised effect estimates and repeated testing of intervention components, and meta-analysis offers a way to update the evidence picture as new studies accrue. The current platform already meets one necessary condition for that ambition by standardising inputs and generating comparable outputs. But the wider methodological literature also warns that early cumulative evidence is unstable: large early effects often attenuate as more trials are added (Ioannidis, 2008; Pereira *et al*., 2012; Cheung and Slavin, 2016; Kraft, 2020; Slavin, 2020). For that reason, the real promise of this model lies in the prospect of a growing series of comparable trials conducted with adequate metadata, common outcomes, and design information that allow heterogeneity to be modelled explicitly. If replication thresholds are met, teacher-led micro-RCTs could become a valuable complement to

large-scale trials: more local, faster, more iterative, and capable of informing both school improvement and broader evidence synthesis.

Implications for development and scale-up

The primary purpose of this proof-of-concept study was to determine whether teachers could successfully use the *WhatWorked Teachers* platform to conduct classroom micro-randomised controlled trials under routine school conditions. Overall, the findings suggest that the platform is operationally feasible, but they also identify several areas where further development is required before wider implementation. These findings should be viewed as informing the next phase of platform refinement rather than representing a final implementation model.

Platform refinements

Several practical improvements were identified through teacher feedback. The onboarding process could be simplified to support more confident selection of trial designs and reduce the creation of duplicate accounts. The workflow has already been modified so that baseline assessment is completed before randomisation, bringing the platform into closer alignment with conventional experimental practice and reducing the potential for allocation awareness to influence pre-test administration. Teachers also highlighted opportunities to reduce administrative burden associated with data entry. While the current Excel-based system provides an important safeguard by ensuring that pupil-identifiable information remains outside the platform, future development should explore ways of streamlining this process without compromising data security. The study also identified opportunities to improve implementation monitoring. Requiring completion of the implementation process evaluation before teachers access impact reports is likely to increase the completeness of implementation data and strengthen future interpretation of trial findings by allowing outcomes to be considered alongside measures of fidelity.

Requirements before wider implementation

Perhaps the most important implication of the study is that further development should focus not only on usability but also on supporting appropriate interpretation and use of evidence. Further work is needed to evaluate alternative approaches to presenting statistical findings so that reports communicate both findings and uncertainty clearly. This should include user testing of Plain English summaries, visual presentation of confidence intervals, and explicit guidance regarding the appropriate use of exploratory findings. The study also highlights important methodological considerations regarding the design of classroom micro-randomised controlled trials. While the platform successfully generated valid statistical analyses, the findings suggest that individual classroom evaluations are unlikely to provide sufficiently precise estimates to support high-stakes instructional decisions in isolation. This reflects the statistical properties of small randomised evaluations rather than a limitation of the analytical platform itself. Consequently, future development should place greater emphasis on helping teachers interpret individual trial findings as one source of evidence alongside professional judgement, existing research evidence and local contextual knowledge, rather than as definitive evidence of intervention effectiveness. Similarly, maintaining clear separation between intervention and comparison conditions emerged as a greater challenge for some trial designs than others. Rather than simply imposing minimum sample sizes, future development should focus on helping teachers select interventions and evaluation designs that maximise treatment contrast and minimise contamination within the practical constraints of classroom delivery.

Future opportunities

Although the present study was not designed to evaluate long-term implementation or system-wide impact, it demonstrates the feasibility of a platform capable of supporting repeated teacher-led evaluations using standardised analytical methods. If future research demonstrates that the methodological challenges identified in this feasibility study can be addressed, the platform could provide infrastructure for accumulating evidence across multiple replications rather than relying solely on findings from individual classroom evaluations. Such cumulative evidence generation may prove particularly valuable in areas characterised by rapid educational innovation, such as educational technology, where conventional large-scale evaluations are often unable to keep pace with the rate at which new products and practices emerge. Similarly, repeated implementation of comparable evaluations across multiple schools could potentially support

future research examining how intervention effects vary across different educational contexts and contribute to understanding which features of interventions appear most consistently associated with positive outcomes. These possibilities remain prospective and require further empirical evaluation. In particular, future research should examine the statistical approaches required to synthesise evidence across repeated classroom evaluations, establish appropriate safeguards for interpreting cumulative findings, and evaluate the costs and resource implications of this model compared with more conventional approaches to educational evaluation.

Relatedly, the platform could complement the EEF's guidance reports by enabling teachers to test recommended approaches within their own contexts. Rather than relying solely on generalised evidence summaries, schools could use the platform to conduct small, structured trials of specific literacy, feedback, or metacognitive strategies and receive rapid feedback on their local impact. Over time, repeated trials across schools could contribute to a cumulative evidence base within particular domains, helping to identify which approaches appear most consistently effective across different contexts. In this sense, the platform has potential not only to support individual teacher enquiry, but also to generate practice-based evidence through the aggregation of many small experiments conducted across diverse school settings. In this way, another possible area for future development concerns contextual interpretation of findings. School and classroom characteristics recorded through the platform could potentially support a 'schools like mine' functionality, enabling teachers to explore how interventions have performed in settings with similar contextual characteristics. This could operate at both national and international levels, which may be particularly valuable in areas where international evidence relating to specific Toolkit strands remains limited for international partners. Such contextualised evidence could help practitioners make more informed judgements about the likely relevance, transferability, and adaptability of findings within their own classrooms and school settings.

Conclusion

Table 1: Summary of findings

Area of research	Findings
Recruitment and engagement	There were 58 schools that expressed an interest in the project, with 19 signing up to take part (33% conversion rate). Teacher accounts were created in 15 schools, and 42 teacher accounts were created, 29 of which were judged to be legitimate teacher accounts (not duplicates or senior managers). Then 25 teachers selected a trial, and 23 completed the randomisation stage. Teachers reported some uncertainty around the trial selection part of the process. The front-loaded spelling approach and class-level randomisation were the more popular options, chosen by 15/18 teachers (83%) and 10/18 (56%) respectively. Then 11/18 teachers completed the teacher surveys and 6 participated in case study interviews.

Trial implementation and attrition	Teachers who progressed to implementation (18/29, 62%) were able to deliver the trial procedures with all 18/18 (100%) of those who administered the pre-tests also completing the post-tests. Front-loaded trials randomised 420 pupils and retained post-test data for 413 pupils (98.3%), while incremental trials randomised 112 pupils and retained post-test data for 93 pupils (83.0%). Teacher confidence with using the platform was high. Platform guidance, assessment administration, and data uploads were reported to be manageable, and the six-week implementation period fitted well within a school half-term. Implementation was more consistent for whole-class trials than for small-group trials, where staffing pressures, timetabling, and maintaining separation between intervention and comparison groups presented greater challenges. The potential risk for contamination was also less likely for whole-class compared to small-group trials.
Interpretation of data	The platform successfully generated standardised statistical analyses and automated impact reports following submission of post-test data. Teachers had some difficulties with interpreting the statistical content in the impact reports, especially effect sizes, confidence intervals, and the meaning of statistical uncertainty. However, some teachers reported discussing findings with colleagues and senior leaders, suggesting potential to stimulate evidence-informed professional dialogue. Due to the different sample sizes in the trials of the front-loaded approach and the low number of trials of the incremental approach, the findings do not provide sufficient evidence to establish a typical effect size and should be interpreted as exploratory. The platform was able to conduct an exploratory meta-analysis, combining the data from the 15 front-loaded trials to explore overall patterns in the findings.

This proof-of-concept feasibility study examined whether teachers could use the *WhatWorked Teachers* platform to conduct classroom micro-randomised controlled trials within routine primary school practice. The findings demonstrate that teachers who progressed to implementation were able to complete the core elements of the evaluation process, including classroom randomisation, intervention delivery, outcome data collection, and automated analysis. Once implementation began, completion of classroom evaluations and outcome data collection was consistently high, indicating that a standardised digital workflow can enable teachers to generate analysable experimental data without requiring specialist research expertise. This represents the principal proof of concept established by the study.

The study also provides a more nuanced understanding of the conditions required for successful implementation. Initial interest from schools was strong, but progression from teacher registration to completed trial was more variable, suggesting that the principal implementation challenge lies in converting early interest into sustained participation. Similarly, while teachers generally understood and accepted randomisation, maintaining clear separation between intervention and comparison conditions proved more challenging for some evaluation designs, particularly within-class small-group trials. The platform therefore appears operationally feasible, but wider implementation will require continued refinement of onboarding, implementation support, and trial design. The findings further demonstrate an important distinction between analytical feasibility and decision-making feasibility. The platform successfully generated automated statistical analyses and impact reports from teacher-entered data, but teachers were less confident interpreting effect sizes, confidence intervals, and statistical uncertainty than they were completing the evaluation process itself. Together with the statistical imprecision inherent in small classroom evaluations, this suggests that findings from individual classroom micro-randomised controlled trials are likely to be most appropriately used to support structured professional enquiry alongside existing research evidence and professional judgement, rather than as a definitive basis for high-stakes instructional decisions.

Importantly, the study was not designed to determine the effectiveness of the spelling interventions or to establish the typical effects achievable through classroom micro-randomised controlled trials. Rather, it demonstrates that teachers can successfully generate analysable experimental data through a standardised digital platform and that this data can be synthesised using common analytical methods. This provides an important methodological contribution by showing that teacher-led experimentation can move beyond informal enquiry towards structured evaluation while also identifying the methodological safeguards needed before wider implementation.

References

Agarwal, P. K., Nunes, L. D., and Blunt, J. R. (2021) 'Retrieval practice consistently benefits student learning: A systematic review of applied research in schools and classrooms'. *Educational Psychology Review*, 33(4), 1409–1453.

Bonell, C., Viner, R. M., Hargreaves, J., Taylor, B., and Ford, T. (2023) 'Approaches to consent in public health research in secondary schools: a narrative review'. *BMJ Open*, 13, e070277.

Breitwieser, J., Neubauer, A. B., Schmiedek, F., and Brod, G. (2024) 'Realizing the potential of mobile interventions for education'. *Science of Learning*, 9(1), 76.

Brown, C. and Greany, T. (2018) 'The evidence-informed school system in England: Where should school leaders be focusing their efforts?' *Leadership and policy in schools*, 17(1), 115–137.

Brown, C. and Zhang, D. (2016) 'Is engaging in evidence-informed practice in education rational? What accounts for discrepancies in teachers' attitudes towards evidence use and actual instances of evidence use in schools?' *British Educational Research Journal*, 42(5), 780–801.

Cain, T. (2015) 'Teachers' engagement with published research: Addressing the knowledge problem'. *The Curriculum Journal*, 26(3), 488–509.

Cantley, I. (2023) 'Replicable quantitative psychological and educational research: Possibility or pipe dream?' *Educational Philosophy and Theory*, 55(1), 111–121.

Cheung, A. C. and Slavin, R. E. (2016) 'How methodological features affect effect sizes in education'. *Educational Researcher*, 45(5), 283–292.

Churches, R. (2016) 'Closing the gap: Test and learn'. *National College for Teaching and Leadership*. **https://assets.publishing.service.gov.uk/media/5a8072e2e5274a2e87db9df2/closing_the_gap_test_and_learn_full_report.pdf**

DEFRA (2025) Statistical digest of rural England. Department for Environment, Food, and Rural Affairs. Available at: https://www.gov.uk/government/statistics/statistical-digest-of-rural-england (Accessed: 23 October 2026).

Department for Education (2022) Key stage 2 attainment: academic year 2021/2022. Updated 15 December. Available at: **https://www.gov.uk/government/statistics/key-stage-2-attainment-provisional-2022** (Accessed: 23 October 2026).

Department for Education (2024a) Key stage 2 attainment: academic year 2022/23. Updated 30 January. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2022-23 (Accessed: 23 October 2026).

Department for Education (2024b) Key stage 2 attainment: academic year 2023/24. Updated 12 December. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2023-24 (Accessed: 23 October 2026).

Department for Education (2025) Schools, pupils and their characteristics: academic year 2024/25. Updated 25 September. Available at: https://explore-education-statistics.service.gov.uk/find-statistics/school-pupils-and-their-characteristics/2024-25 (Accessed: 23 October 2026).

Dymock, S. and Nicholson, T. (2017) 'To what extent does children's spelling improve as a result of learning words with the look, say, cover, write, check, fix strategy compared with phonological spelling strategies?' *Australian Journal of Learning Difficulties*, 22(2), 171–187.

Esposito, R., Herbert, E., and Sumner, E. (2022) 'Capturing variations in how spelling is taught in primary school classrooms in England'. *British Educational Research Journal*, 49(1), 70–92.

Fisher, B., Cozens, M. E., and Greive, C. (2007) 'Look-say-cover-write-say-check and old way/new way—mediational learning: A comparison of the effectiveness of two tutoring programs for children with persistent spelling difficulties', *Special Education Perspectives*, 16(1), pp. 19-38.

Graham, S. and Santangelo, T. (2014) 'Does spelling instruction make students better spellers, readers, and writers? A meta-analytic review'. *Reading and Writing*, 27, pp.1703–1743.

Hargreaves, L. M. (2009) 'Respect and responsibility: Review of research on small rural schools in England'. *International journal of educational research*, 48(2), 117–128.

Hase, A. and Kuhl, P. (2024) 'Teachers' use of data from digital learning platforms for instructional design: a systematic review'. *Educational technology research and development*, 72(4), 1925–1945.

Hedges, L.V. and Rhoads, C. (2010) *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Ioannidis, J. P. (2008) 'Why most discovered true associations are inflated'. *Epidemiology*, 19(5), 640–648.

Klasnja, P., Hekler, E. B., Shiffman, S., Boruvka, A., Almirall, D., Tewari, A., and Murphy, S. A. (2015) 'Microrandomized trials: an experimental design for developing just-in-time adaptive interventions'. *Health Psychology*, 34(S), p.1220.

Kraft, M. A. (2020) 'Interpreting effect sizes of education interventions'. *Educational Researcher*, 49(4), 241–253.

Leat, D., Reid, A., and Lofthouse, R. (2015) 'Teachers' experiences of engagement with and in educational research: What can be learned from teachers' views?' *Oxford review of education*, 41(2), 270–286.

Leckie, G. and Goldstein, H. (2019) 'The importance of adjusting for pupil background in school value-added models: A study of Progress 8 and school accountability in England'. *British Educational Research Journal, 45*(3), 518–537. https://doi.org/10.1002/berj.3511

Levin, B. (2011) 'Mobilising research knowledge in education'. *London Review of Education*, 9(1).

Levin, B. (2013) 'To know is not enough: Research knowledge and its use'. *Review of Education*, 1(1), 2–31.

Liao, P., Klasnja, P., Tewari, A., and Murphy, S. A. (2016) 'Sample size calculations for micro-randomized trials in mHealth'. *Statistics in Medicine*, 35(12), pp.1944–1971.

Liu, C., Cox Jr., R. B., Washburn, I. J., Croff, J. M., and Crethar, H. C. (2017) 'The effects of requiring parental consent for research on adolescents' risk behaviors: a meta-analysis. *Journal of Adolescent Health*,' 61(1), pp.45–52.

Makel, M. C. and Plucker, J. A. (2014) 'Facts are more important than novelty: Replication in the education sciences'. *Educational Researcher*, 43(6), 304–316.

Mawson, R. D. and Kang, S. H. (2025) 'The distributed practice effect on classroom learning: A meta-analytic review of applied research'. *Behavioral Sciences*, 15(6), 771.

McGlade, H. and Kelly, J. (2025) Investigating educational disadvantage and place-based approaches in the North East of England. Education Policy Institute, Durham University, North East Combined Authority. Available at: **https://epi.org.uk/wp-content/uploads/2025/05/EPI-Durham-University-and-NECA_summary-paper_May25_V2.pdf** (Accessed: 23 October 2026).

Mouton, B., Loop, L., Stievenart, M., and Roskam, I. (2018) 'Child differential sensitivity to parental self-efficacy improvement: a micro-trial perspective'. *International Journal of Behavioral Development*, 42(2), pp.203–213.

Nelson, J. and Campbell, C. (2017) 'Evidence-informed practice in education: Meanings and applications'. *Educational Research*, 59(2), 127–135.

OECD (2023) Who really cares about using education research in policy and practice? Developing a culture of research engagement. Available at: **https://www.oecd.org/en/publications/who-really-cares-about-using-education-research-in-policy-and-practice_bc641427-en.html** (Accessed: 23 October 2026).

Ofsted (2024) Telling the story: the English education subject report. Updated 5 March. Available at: https://www.gov.uk/government/publications/subject-report-series-english/telling-the-story-the-english-education-subject-report (Accessed: 23 October 2026).

Owen, K. L., Watkins, R. C., and Hughes, J. C. (2022) 'From evidence-informed to evidence-based: An evidence building framework for education'. *Review of Education*, 10(1), e3342.

Parker, K., Nunns, M., Xiao, Z., Ford, T., and Ukoumunne, O. C. (2021) 'Characteristics and practices of school-based cluster randomised controlled trials for improving health outcomes in pupils in the United Kingdom: a methodological systematic review'. *BMC Medical Research Methodology*, 21(1), 152.

Pereira, T. V., Horwitz, R. I., and Ioannidis, J. P. (2012) 'Empirical evaluation of very large treatment effects of medical interventions'. *Jama*, 308(16), 1676–1684.

Perry, T., Morris, R., and Lea, R. (2022) 'A decade of replication study in education? A mapping review (2011–2020)'. *Educational research and evaluation*, 27(1–2), 12–34.

Preston, D., Qualter, P., and Woods, K. (2026) 'A systematic review of facilitators and barriers to school staff development projects using the Consolidated Framework for Implementation Research'. *British Educational Research Journal*.

Prior, L., Jerrim, J., Thomson, D., and Leckie, G. (2021) 'A review and evaluation of secondary school accountability in England: Statistical strengths, weaknesses, and challenges for 'Progress 8' raised by COVID-19'. *Review of Education, 9*(3), e3299. https://doi.org/10.1002/rev3.3299

Qian, T., Walton, A. E., Collins, L. M., Klasnja, P., Lanza, S. T., Nahum-Shani, I., and Murphy, S. A. (2022) 'The microrandomized trial for developing digital interventions: Experimental design and data analysis considerations'. *Psychological methods*, 27(5), 874.

Rickinson, M., Walsh, L., de Bruin, K., and Hall, M. (2019) 'Understanding evidence use within education policy: A policy narrative perspective'. *Evidence and Policy*, 15(2), 235–252.

Sharples, J., Eaton, J., and Boughelaf, J. (2024) A School's Guide to Implementation. Education Endowment Foundation: London. Available at: https://educationendowmentfoundation.org.uk/education-evidence/guidance-reports/implementation (Accessed: 23 October 2026).

Shaw, T., Cross, D., Thomas, L. T., and Zubrick, S. R. (2015) 'Bias in student survey findings from active parental consent procedures'. *British Educational Research Journal*, 41(2), pp.229–243.

Sim, J. (2019) 'Should treatment effects be estimated in pilot and feasibility studies?' *Pilot and feasibility studies*, 5(1), 107.

Slavin, R. E. (2020) 'How evidence-based reform will transform research and practice in education'. *Educational Psychologist*, 55(1), 21–31.

Topping, K. J., Thurston, A., McGavock, K., and Conlin, N. (2012) 'Outcomes and process in reading tutoring'. *Educational Research*, 54(3), pp.239–258.

UK Parliament, Parliamentary Office of Science and Technology. (2024, November 11) Education: Inequalities and attainment gaps. POST. Available at: **https://post.parliament.uk/education-inequalities-and-attainment-gaps/** (Accessed: 23 October 2026).

WhatWorked Education (2025) Building a school-led evidence base in the North East: final report. September. Available at: https://interventions.whatworked.education/school-improvement-2026 (Accessed: 23 October 2026).

Appendices:

Appendix A – K can be found published in a separate document on the project page.

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

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

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

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



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