Cognitive science Teacher Choices trial: Using examples to teach grammar to Year 7

Evaluation report

October 2025

Katherine Aston, Andrew Smith, Jose Liht, Eleanor Bradley, Rob Ager, Michael Thomas, Andrew Tolmie, Annabel Watson and Helen Poet

The Education Endowment Foundation (EEF) 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:

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

0207 802 1653

info@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator

The project was independently evaluated by a team from the National Foundation for Educational Research (NFER), in partnership with Birkbeck, University of London, University College London (UCL), and the University of Exeter. The trial director and principal investigator for this study was Helen Poet, Research Director. Katherine Aston, Research Manager, led the evaluation team and the implementation and process evaluation (IPE), supported by Eleanor Bradley, IPE Researcher. Andrew Smith led the impact evaluation, with Jose Liht and Gemma Schwendel as trial statisticians. Kathryn Hurd led the research operations, with Katharine Stoodley, Lydia Wallis, Sarah Millar, and Holly Critchley as operations researchers. Rob Ager, Frances Brill, and Sarah Gibb led the assessment development and marking, supported by Katharine Larkin.

Professor Michael Thomas (Birkbeck) and Professor Andy Tolmie (UCL) advised on cognitive science and led the initial literature review. Dr Annabel Watson (Exeter) advised on English pedagogy and led the development of the teacher guidance and resources.

Contact details:

Katherine Aston
National Foundation for Educational Research (NFER)
The Mere,
Upton Park,
Slough,
Berkshire,
SL1 2DQ

Tel: 01753 574123
Email: k.aston@nfer.ac.uk

Acknowledgements

We are grateful to all the secondary schools that participated in this evaluation. We are especially thankful to the staff and pupils who took part in the teaching and those who talked to us during the interviews and focus groups, as well as those who responded to the surveys and completed the assessments. We would like to thank Bob Pritchard, Niki Kaiser, Gaia Scerif, Steve Higgins, Joshua Clarke, and Lisa-Maria Müller who served as members of the Study Advisory Board.

We would like to thank the team at the Education Endowment Foundation, including Daniela Alvarado, Lauren Spinner, Toby Whittaker, Igraine Rhodes, and Rob Coe.

Executive summary

The project

This project evaluated the impact of three approaches to using worked or non-worked examples to teach grammar in Year 7 English. It aims to extend research on cognitive science and worked examples to Key Stage 3 English, in the context of 'real-world' classroom use. This is the Education Endowment Foundation (EEF) Teacher Choices trial, which look to test everyday choices teachers make when planning their lessons and supporting their pupils.

Two of the approaches tested were 'worked example' approaches, where teachers first model how a grammar pattern is constructed, before pupils write their own text using the pattern. These were:

1. **Systematic use of worked examples (systematic worked arm).** Around 20 sessions, spaced twice a week as a separate activity in the lesson.
2. **Responsive use of worked examples (responsive worked arm).** Around 20 sessions, taught based on the teacher's judgement of where the activity fits into the curriculum for each half-term block.

The 'worked examples' approaches were contrasted with:

3. **Non-worked examples (non-worked arm).** Around 20 sessions where teachers were asked to discuss the grammar pattern in the examples, and the effect of the grammar pattern on the reader. Teachers were asked not to use the examples as a model for pupils' own writing.

Teachers were sent brief written guidance for using their allocated approach and an example session plan. They were sent a bank of examples, which they could use or adapt as they preferred.

The project was a three-armed, cluster randomised controlled trial. Year 7 English teachers and pupils (aged 11–12) in 55 state-funded secondary schools in England were allocated to test one approach for ten weeks in one term. Teachers were asked to teach 20 (15-minute) sessions during the trial, totalling five hours of teaching. Within each school, the National Foundation for Educational Research (NFER) grouped teachers and classes into small 'teacher-class units' to ensure that each teacher and class was only allocated to one approach.

The trial's primary outcome was a bespoke assessment of writing composition, aligned with the two writing genres covered in the grammar sessions. The implementation and process evaluation (IPE) involved surveys, pupil focus groups, teacher interviews, and observations. It explored choice implementation, the role of teacher guidance, and perceived engagement and outcomes. The trial was designed and evaluated by NFER, with academic advisors from Birkbeck, University of London, University College London (UCL), and the University of Exeter.

Table 1: Key conclusions

Key conclusions	
1.	There was no evidence of meaningful differences between approaches to using examples on pupils' writing assessment scores.
2.	There was no evidence of meaningful differences between the approaches to using examples on the writing assessment scores of pupils eligible for free school meals (FSM).
3.	There was no evidence that prior attainment influenced the effect of different teaching approaches on pupils' writing assessment scores.
4.	The teaching approaches represented a substantial change to usual practice for many teachers. Teachers reported that a sustained focus on grammar patterns within text was new to their teaching and their classes, particularly elements of worked examples, such as modelling the step-by-step construction of a grammar pattern or asking pupils to follow that step-by-step construction in their writing. Given this substantial change, additional support for teachers may have been needed to achieve sufficient contrast between the approaches.

5. Teachers in the worked example approaches perceived that most pupils could successfully use a grammar pattern in their writing when this was highly scaffolded. However, teachers perceived that pupils rarely transferred use of the taught grammar patterns into more general writing composition tasks and suggested pupils would need additional support to do so.

Security of the trial

This was a Teacher Choices trial, testing an everyday choice teachers make when teaching grammar. The trial was a well-designed, three-armed, randomised controlled trial and was well powered. The pupils were similar across the three arms. While the missing data for the primary outcome (29%) reduced the security of the trial findings, the missing data was evenly distributed across the three arms, reducing the risk of bias and supporting the reliability of the results.

Additional findings

Writing assessment scores for all pupils, presented as adjusted mean scores, were similar across the three teaching approaches. These averaged approximately 18 out of 40 marks for all three approaches, as shown in Table 2 below. The small differences in the adjusted mean scores and overlapping confidence intervals (CIs) do not provide evidence of meaningful differences between approaches for all pupils or for the subsample of FSM-eligible pupils. These findings suggest that there was no evidence of different impacts on pupil writing for the different teaching approaches, contrary to what was anticipated at the start of the trial, which was that worked examples would have a more positive impact. Confidence in this finding is somewhat limited by the high percentage of missing data (29% of all randomised pupils) and should be interpreted alongside the considerations mentioned in the 'Security of the trial' section above. However, it reflects the findings from other writing trials (e.g. Torgerson *et al.*, 2018; Anders *et al.*, 2021), which show the challenges of having an impact on writing.

As each of the teaching approaches represented a significant change in practice for teachers, it is unclear how they compared with teachers' usual practice. The approaches could have had a positive effect, negative effect, or no impact. Challenges in implementation, as well as challenges in the transfer of learning, may have contributed to this result. First, the three example approaches were taught more similarly than intended, reducing the choice contrast. Second, teachers reported that while the worked example approaches enabled most pupils to construct a text, which included the grammar pattern, pupils rarely independently transferred the grammar patterns into other writing tasks. This would explain the absence of an effect on the primary outcome of writing composition. Teachers did not adapt the approaches and content taught as much as we expected, which appeared to be due to a combination of time/workload pressures and a perception that they needed to always use the optional example grammar patterns and model texts.

Overall, teachers responded positively to the three example approaches, were able to implement them, and felt they met a need to develop grammar teaching in Key Stage 3 English. They reported gains in their knowledge for teaching grammar, and interviewed teachers intended to embed the worked examples approaches in their future teaching. Pupils were positive about the worked example approaches. This shows that it is feasible to use worked examples in areas of the curriculum beyond maths and science.

Impact

Table 2: Summary of impact on primary outcome

Outcome / group	Examples	Adjusted means (95% CI)	Standard error	No. of pupils
Text-type Specific Writing Assessment (TSWA) – all pupils	Systematic worked	18.17 (17.66, 18.67)	0.26	2,022
	Responsive worked	17.92 (17.44, 18.41)	0.24	2,408
	Non-worked	17.60 (17.11, 18.08)	0.25	1,867
TSWA – FSM-eligible pupils only	Systematic worked	16.89 (16.18, 17.59)	0.36	421
	Responsive worked	16.23 (15.59, 16.86)	0.32	646
	Non-worked	15.89 (15.25, 16.53)	0.33	473

Introduction

Background

This study is a 'Teacher Choices' evaluation focused on the use of cognitive science in the classroom. It looks at different uses of examples to teach grammar in Key Stage 3 English lessons.

Teacher Choices trials

Teacher Choices trials explore some of the most common questions teachers ask about their practice and the everyday choices they make when planning lessons and supporting pupils. The aim of Teacher Choices research is to investigate the impact of these different day-to-day pedagogical practices on pupil learning and to generate evidence that can be readily applied by teachers in the classroom. This is a new and developing strand of the Education Endowment Foundation (EEF)-funded research. Teacher Choices trials aim to explore choices, which are of high interest to schools, a real choice that can be made by classroom teachers, and easy to implement without intensive training and resources.

Cognitive science and using examples

Within their Teacher Choices programme, the EEF selected cognitive science as a trial topic because it is an area with a good theoretical and experimental evidence base, with potential to have a differential effect on those from socio-economically disadvantaged backgrounds (Perry *et al.*, 2021). It is particularly suited to a Teacher Choices trial because there is currently less research based on day-to-day classroom practice.

In early 2023, the EEF commissioned the National Foundation for Educational Research (NFER) to undertake scoping work for a possible trial at Key Stage 3 exploring: 'Which modelling technique that uses examples is most effective?' This research question had been selected based on high interest from secondary teachers in a preliminary survey conducted by the EEF using the teacher survey panel Teacher Tapp, to understand teachers' priorities for research in cognitive science in the classroom.

At the start of scoping, we refined the research question to: 'What approach to the use of worked examples is most effective?' Worked examples are a form of modelling, which provides a step-by-step demonstration and describes the process of completing the task. We focused the question on worked examples because they have a strong basis in cognitive science research. They are expected to support learners' formation of schemas, which are the automated mental frameworks used to organise, process, and store information (Bartlett, 1932; Whitney, 2001). These schemas serve to reduce cognitive load, which is the mental processing capacity required to manage task demands (Sweller, 1994).

Scoping work

Scoping work comprised of:

- a review of the literature on worked examples;
- teacher consultation via interviews;
- a Teacher Tapp survey; and
- two teacher co-design sessions, which brought together teachers, evaluators, and cognitive science specialists to identify contrasting choices and give feedback on trial design options (e.g. preferred level of randomisation, trial length, and the feasibility of changing practice).

A review of current literature (67 in-scope sources, comprising 2 meta-analyses and 65 empirical reports) was conducted in early 2023 to establish what research had been carried out on the use and timing of worked examples, both by researchers in laboratory settings and by teachers/researchers in classroom contexts. The review considered the age groups employed,

the subject contexts used, the intervention length and design factors explored, the assessment of impact, and the outcomes reported. The review is included in the project study plan (Smith *et al.,* 2024).
The key conclusions were as follows:

Overall, the evidence summarised in the literature review strongly favoured positive effects of using worked examples, although there was little work in the context of teaching English. However, the consistency of outcomes across maths and science suggested that positive effects might reasonably be expected in other subject contexts. Age did not appear to be a material factor either. The review suggested that the use of worked examples was probably better when pupils work individually rather than collaboratively. The most effective sequence of use appeared to be overarching examples then individual engagement with the detail of these, with backward fading (gradual removal of explicit steps) over successive stages of activity. Light-touch support and feedback from teachers may increase the impact of worked example use. The robustness over time of these effects was essentially unknown, because of the lack of anything beyond a minimal delay in post-testing. The immediate effects reported were predominantly based on single-session interventions, and it may be reasonable to expect that longer interventions, over a matter of weeks, would be required to achieve lasting effects (see e.g. Thurston *et al.,* 2009). To detect robust effects, there was a need for both proximal (i.e. near-immediate, topic-specific) and distal testing, with the latter focused on more general tests if suitable measures that could plausibly be influenced by the worked examples could be identified. Our trial design addressed these points by using a ten-week intervention period and measuring overall writing composition as the primary outcome.

After completing the review, the EEF and the evaluation team, in consultation with the Study Advisory Board, decided to focus the scoping phase on English, as it was deemed feasible and had the potential to extend both understanding and classroom practice in important ways. In addition, there was evidence of substantial interest in applying the principles of cognitive science in teaching English. A Teacher Tapp survey commissioned by the EEF showed that secondary English teachers had a high level of interest in answering the research question: 'Which modelling technique that uses examples is most effective?' Consultations with Key Stage 3 English teachers during the scoping phase showed that they use examples throughout their teaching, including worked examples, with practice varying significantly. The Office for Standards in Education, Children's Services and Skills (Ofsted) (2022) research review for English recommended using worked examples, particularly with novice learners, to draw attention to specific features of writing and to reduce cognitive load when pupils undertake complex tasks (Kyun, Kalyuga, and Sweller, 2013; Graham, Harris, and Chambers, 2016). However, there was insufficient subject-specific evidence on the effectiveness of worked examples to provide evidence-informed advice on using worked examples in English.

This trial therefore, aimed to extend existing knowledge about the impact of worked examples beyond application to just maths and science, and in realistic classroom contexts. This aligned with teacher research priorities for cognitive science, collected via a survey and analysed by the Chartered College of Teaching, which highlighted teachers' need to understand how cognitive science can be implemented effectively in specific phases and subjects (Müller and Cook, 2023).

Developing the Teacher Choices

Developing the choice of teaching approaches to be tested required integrating the cognitive science principles of worked examples into the disciplinary and teaching context of English, and specifically the chosen trial subject content of grammar. Grammatical knowledge in the secondary English curriculum represents an 'ill-structured' domain (Kyun, Kalyuga, and Sweller, 2013, p. 386). While explicit knowledge of grammatical constructions and terminology is tested at the end of Key Stage 2, subsequent grammatical knowledge is assessed through application in reading and writing activities, with pupils required to study 'the effectiveness and impact of the grammatical features of the texts they read' (DfE, 2013, p. 5) and to draw on 'grammatical constructions from their reading and listening' (ibid, p. 5) and use these 'consciously in their writing and speech to achieve particular effects' (ibid, p. 5). This approach is supported by a body of research, which embeds the teaching of grammar in the context of reading and writing with the goal of expanding pupils' metalinguistic understanding of the linguistic choices made in writing (Jones, Myhill, and Bailey, 2013; Myhill and Watson, 2014; Chen and Myhill, 2016). This requires an interweaving of declarative and procedural knowledge: the explicit teaching of knowledge about grammar aims to expand the repertoire of sophisticated grammatical choices that pupils can make in their writing, alongside their ability to identify and analyse the choices that writers make in reading activities.

While writing is commonly conceptualised as a problem-solving activity (e.g. Kellogg, 1999), a given writing 'problem' usually has a huge range of potential successful solutions and numerous different ways to tackle it: in cognitive terms, the 'goal-state' is poorly specified and the 'problem solving operators are unspecified' (Kyun, Kalyuga, and Sweller, 2013, p. 386). Nevertheless, use of authentic text models—particularly when accompanied by guided analysis and discussion, which links the form used in the model to the way in which it communicates meaning to the reader—is frequently advocated as an effective way to support writing development (e.g. Myhill, Lines, and Jones, 2018; Graham *et al.*, 2016). These models are rarely referred to as 'worked examples', though there are cases of the term being used synonymously with 'text models' in writing instruction (e.g. McLoughlin, 2008). While Kyun, Kalyuga, and Sweller (2013) operationalised worked examples simply as example essays provided to writers, the emphasis on the role of discussion and analysis in grammar research (e.g. Newman and Watson, 2020) suggests that additional scaffolding in the form of guided analysis and mimesis may be particularly helpful for supporting the movement between declarative knowledge of grammatical forms and procedural application in writing activities, as pupils combine analysis of focus constructions with opportunities to experiment with crafting new examples (Watson, Newman, and Morgan, 2021). Backward fading in this context involves the gradual reduction and removal of this teacher guidance. It has been suggested that the benefits of such guided analysis and imitation activities may not just be to teach pupils to use the particular grammatical form that they are examining, but rather that the activities help to develop pupils' sensitivity to language choices more generally: 'There is a strong argument that the value of texts as models is less that they offer models for imitation but that they open up metalinguistic awareness of the repertoire of possibilities of language choices' (Myhill, Lines, and Jones, 2018, p. 8).

Cognitive science theorises that worked examples assist the process of schema formation and consolidation with respect to the tasks to which those examples are applied, by: i) providing a clear mapping of how to approach the task; ii) repeated performance, with active retrieval as steps are removed; and iii) reducing cognitive load and therefore, permitting greater focus on task structure (Tarmizi and Sweller, 1988; Chen, Kalyuga, and Sweller, 2015). In both worked example conditions, teacher-guided analysis of authentic examples was followed by independent pupil review of further examples of the same grammatical pattern, and then by guided imitation in pupil writing (cf. Reiss *et al.*, 2008). The examples were expected to facilitate the transfer of declarative knowledge of grammatical forms and their impact on the reader into procedural ability to consciously and purposefully deploy these patterns in pupils' own writing (cf. Atkinson, Renkl, and Merrill, 2003). As a consequence, pupils were expected to show increased use of the specific forms taught (cf. Richey and Nokes-Malach, 2013; McLaren *et al.*, 2016) but also to become more generally sensitive to the ways in which they can manipulate grammar to communicate with the reader (cf. Van Gog, Paas, and Van Merriënboer, 2006), benefiting wider performance in writing.

The trial had two worked example arms, systematic and responsive, which reflect two common approaches to teaching grammar in secondary English, based on the Teacher Tapp survey from our scoping phase (see Appendix P). Systematic use entailed use of worked examples as focused lesson starters regularly throughout the trial period. Responsive use of worked examples asked teachers to identify planned or spontaneous opportunities to use worked examples for grammar patterns within their existing teaching scheme. These were intended to be integrated into the lessons and may not have been regularly spaced across the trial. As teaching schemes vary widely, we expected greater heterogeneity in practice within the responsive worked arm.

As well as reflecting different approaches to teaching and structuring lessons, the rationale for including two different delivery patterns of worked examples (i.e. systematic and responsive) reflected the different perspectives of cognitive science and English pedagogy.

Cognitive science literature on worked examples (cf. Barbieri *et al.*, 2023), suggests that systematic use should lead to schema development and improvement in grammar and writing outcomes across all the content covered. Here, improvement was hypothesised to be consistent across the class, that is, that all pupils in the class will see improvement in the stated outcomes, including pupils previously performing less well. As the responsive worked arm does not feature separate starters focused on worked examples, and use of worked examples may not be spaced regularly through the trial period, cognitive science (e.g. Perry *et al.*, 2021) suggests it would have a smaller and less consistent effect on pupil outcomes.

In contrast, English pedagogical literature on contextualised grammar (e.g. Jones, Myhill, and Bailey, 2013; Myhill, Jones, and Lines, 2018) suggests that responsive use of worked examples would show a larger improvement in grammar and writing outcomes than systematic use. This hypothesis is because using worked examples responsively integrates them within the broader curriculum context and focuses on diagnosed pupil needs. In this theory, systematic use of worked examples is hypothesised to have a smaller effect than responsive use, since it risks artificially separating grammar from the broader learning of pupils, potentially reducing transfer, and paying equal attention to all grammatical constructions irrespective of pupil need or progress.

Use of non-worked examples is hypothesised to lead to unsystematic individual improvement in grammar and writing outcomes, since the lack of direct support for schema formation means that only those capable of extracting task features for themselves will benefit. This condition was expected to support declarative knowledge of grammatical constructs (the secondary outcome) and the ability to explain their potential impact on the reader. However, given that there is no support for the transfer of declarative knowledge into procedural application in original writing, improvement in the application of this knowledge to writing was expected to be individual and unsystematic.

None of the three approaches was conceptualised as business as usual across schools. Instead, Teacher Choices trials aim to compare active contrasting choices, each of which may be business as usual for some teachers and/or schools. Each of the three approaches to using examples is grounded in practices described by teachers in our scoping phase, although we did not assess their prevalence in practice across schools. Teachers indicated that they used a variety of different approaches to using examples within their practice. By comparing the three approaches with each other, we aimed to improve our understanding of patterns of use of examples that are most effective in helping pupils develop more sophisticated grammar choices in their writing.

In the two 'worked examples' approaches tested in the trial, teachers were asked to use model texts to identify steps to build the grammar construction. Teachers were asked to guide pupils to work through these steps to develop their own text, which uses the pattern. In the systematic worked approach, these teaching episodes were intended to be separate lesson starters, regularly spaced throughout the trial. In the responsive worked approach, teachers were asked to integrate the same number of teaching episodes into lessons within the teaching scheme.

In the 'non-worked' examples approach, teachers were asked to guide analysis of model texts to understand the grammar construction and explore its effect. In contrast to the 'worked' examples, pupils were not asked to 'work' through steps to construct their own written examples. These teaching episodes were intended as separate lesson starters, regularly spaced throughout the trial.

Our implementation and process evaluation (IPE) surveys explored the prevalence of these elements in teachers' usual practice, and the ease of implementing the allocated approach, in order to understand choice differentiation. These findings are reported in the 'IPE results' section below.

At the end of the scoping phase (July 2023), NFER and the EEF agreed to conduct a trial based on the three approaches to using examples to teach grammar for writing. Teacher guidance was developed for each of the three approaches. Options discussed including written guidance and exemplars, videos of practice, webinars, and ad hoc email support. As Teacher Choices are intended to operate with minimal guidance, it was decided that teachers would receive written guidance to follow their approach, including an example session plan. All teachers also received an optional bank of examples for ten grammar patterns, designed for use with any of the approaches.

Teacher Choice approaches

The following Template for Intervention Description and Replication (TIDieR) framework (Table 3) outlines the choices, which were tested in this trial (for further details also see the project study plan; Smith *et al.*, 2024). This description sets out the intended ideal practice for each teaching approach. Teacher adherence and fidelity to these idealised practices are discussed in the 'IPE results' section below.

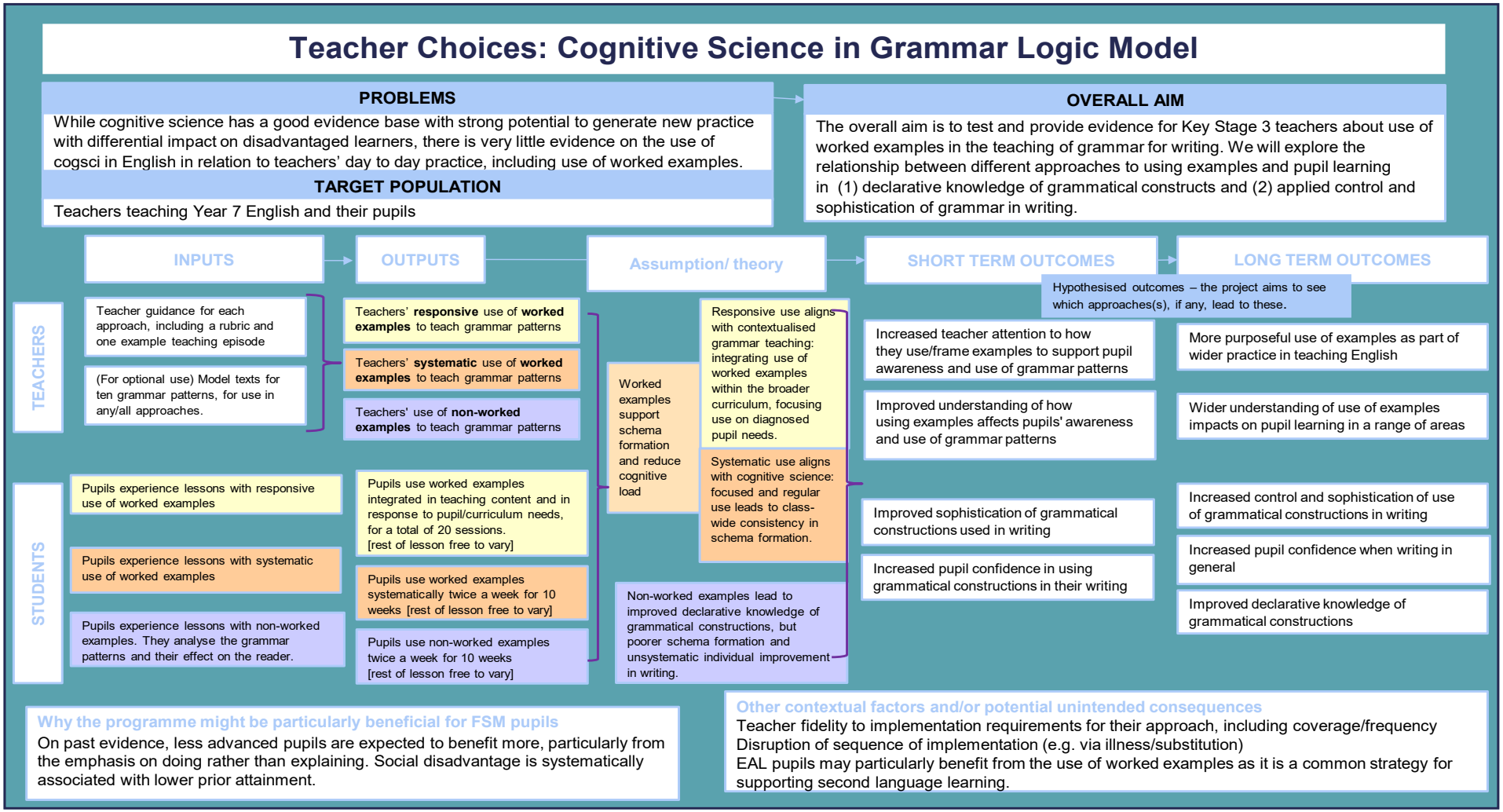
Figure 1 below shows the logic model for the three choices.

Table 3: Intervention description (TIDieR)

Name of example approach		Systematic worked	Responsive worked	Non-worked
Why (rationale)		**Worked examples are expected to support schema formation and reduce cognitive load. Systematic use aligns with cognitive science: focused and regular use is expected to lead to class-wide consistency in schema formation**	**Worked examples are expected to support schema formation and reduce cognitive load. Responsive use aligns with contextualised grammar teaching: integrating use of worked examples within the broader curriculum, focusing use on diagnosed pupil needs**	**Non-worked examples are expected to lead to improved declarative knowledge of grammatical constructions, but poorer schema formation and unsystematic individual improvement in writing**
Who (practitioners and recipients)		Teachers of Year 7 English were asked to use the teaching approaches. Year 7 pupils from participating schools received the teaching approaches. Teachers were asked to use examples with all pupils in their class, including pupils with Special Educational Needs and Disabilities (SEND) and pupils with English as an Additional Language (EAL)		
What (teaching)	Content	Teachers were asked to teach grammatical constructions for 'noun phrases' in the first half-term block, and 'clause/sentence' constructions in the second half-term block. Content and sequencing were flexible within each half-term block		
	Level of integration	Separate teaching episode e.g. lesson starter	Embedded at any point in the lesson, to be determined by the teacher	Separate teaching episode e.g. lesson starter
	Grammar patterns	One pattern in each week, taught twice	Flexible as needed	One pattern in each week, taught twice
	Total no. of sessions	20	20	20
	Frequency of sessions	Twice a week in different English lessons (required)	Flexible within the trial period, (completing 20 sessions across the ten weeks)	Twice a week in different English lessons (required)
	Focus of input	The teacher uses model texts to identify steps to build the grammatical construction Pupils work through these steps to develop their own text, which uses the pattern	The teacher uses model texts to identify steps to build the grammatical construction Pupils work through these steps to develop their own text, which uses the pattern	The teacher guides the analysis of model texts to understand the grammatical construction and explore its effect No pupil writing to construct their own examples

What (teacher guidance materials)	A short guide with 'dos and don'ts' to enable them to implement their allocated approach, with one example episode plan (15-minute session). An optional example bank, comprising five grammatical constructions for each topic (e.g. pre-modifying nouns with adjectives) with two sets of examples for each focus form. Each set of examples comprised one authentic text model (from a story or speech) and three examples constructed to follow the same linguistic pattern Participating teachers only received the guidance for their own allocated approach and were asked not to discuss the approaches with teachers in other arms, to reduce the risk of contamination. Trial leads had access to all three sets of guidance The teacher guidance materials and optional example bank are included in Appendix F and Appendix G
Where (location)	Classrooms in state-funded secondary schools in England
When and how much (dosage)	For 20 sessions of 15 minutes in Summer Term 2024
Tailoring (adaptation)	To participate in the trial, teachers were asked to teach grammar patterns based on the content outlined above, to teach 20 sessions in total, keeping their session length close to 15 minutes. They were asked to read and follow the implementation guidance document for the allocated approach and make a credible attempt to comply with their approach Teachers could choose: • which grammar patterns to teach for noun phrases in narrative fiction and clauses/sentences in persuasive speeches; and • whether to use or adapt the grammar patterns and model texts provided, or develop their own examples To minimise contamination, teachers using the non-worked examples were instructed not to ask pupils to imitate the focus constructions in their own writing, either during the episode or in the remainder of any lessons taught during the trial period. Pupils could do this independently if they were undertaking writing activities, but it would not form part of teaching

Figure 1: Logic model (as agreed in the project study plan)



Project delivery

We recruited fewer schools than expected (55 compared with 65). Reasons for this included a relatively short timeline for recruitment (informal Expressions of Interest from November 2023 to December 2023, and formal recruitment from January 2024 to mid-March 2024) and a low response rate to recruitment materials emphasising cognitive science in English. To mitigate for this, we developed further recruitment materials emphasising the development of grammar for writing, from which we recruited additional schools.

We also had higher attrition than expected (29.3% pupil-level attrition in the primary analysis). The main reasons for attrition were schools or classes withdrawing from the evaluation before testing (15 of the 222 teacher-class units, five from each allocated teaching approach), or because the writing assessment was returned to NFER unused, most commonly due to pupil absence on the day of testing, or whole classes not completing the assessment. Overall, this did not impact the trial's planned statistical power, as the actual school and class intracluster correlation coefficients (ICCs) were lower than expected, and the pre- and post-test correlations were higher than expected. There is some evidence from the IPE that the implementation of different teaching approaches was less distinct than intended. These points are discussed further in the 'Methods', 'Impact evaluation results' and 'Implementation and process evaluation' sections below.

Overview of evaluation design

This evaluation is a randomised controlled trial, which aimed to measure the impact of three different approaches to using examples to teach Year 7 grammar on pupil writing composition.

The research design is a three-arm trial with randomisation of teacher-class units[1] (for further details see 'Methods' section under 'Trial design' subsection below, p. 20). Each teacher-class unit was randomly allocated to one grammar teaching approach to use throughout one term (Summer Term 2024) and asked to implement it for each of their Year 7 English classes.

Evaluation objectives

The evaluation objectives (Table 4) and research questions were pre-specified and published in the project study plan (Smith *et al.,* 2024), and summarised below.

Table 4: Overview of the evaluation objectives and research questions

Research objective	Research questions	Methodology area	Data collection methods and participants	Data analysis methods
Compare the impact of example approaches on writing composition	1	Impact evaluation	Pupil assessment and the National Pupil Database (NPD) secondary data	Intention-to-treat (ITT) analysis[a] Likelihood ratio tests comparing multi-level models, which include the three treatment arms (where appropriate followed by post-hoc pairwise comparisons, and estimation of effect sizes)
Compare the impact of example approaches on knowledge of grammatical constructs	2	Impact evaluation	Pupil assessment and NPD secondary data	
Measure variation in impact by prior attainment	3	Impact evaluation	Pupil assessment and NPD secondary data	
Measure variation in impact by the number of sessions taught	4	Impact evaluation	Teacher dosage log, pupil assessment, and NPD secondary data	
Explore differential effects for FSM pupils	1a, 2a	Impact evaluation	Pupil assessment and NPD secondary data	
Describe the fidelity of implementation and differentiation	5	IPE	Teacher baseline and endpoint surveys, dosage log	Statistical analysis (frequencies, cross-tabulations)
Describe responsiveness	6	IPE		

[1] To ensure that each teacher and each class was assigned to a single approach, in a context with high prevalence of shared teaching across classes, NFER combined teachers and classes into discrete teacher-class units, defined as the smallest possible number of teachers where there was no shared teaching with another teacher-class unit.

Describe perceived outcomes	7	IPE	Case studies (observation, teacher interviews, pupil focus groups)	Qualitative thematic analysis Integration of findings from different methods
Evaluate the feasibility of implementing teaching approaches, including the role of guidance	8	IPE		

[a]*Compliance analysis was not undertaken due to missing data, this is discussed in the 'Impact evaluation' subsection in the 'Methods' section below.*

The primary aim of this trial is to estimate the impact on pupil learning through three different approaches to using examples to teach Year 7 grammar.

The logic model identifies the potential long-term pupil outcomes as:

1. Increased control and sophistication of use of grammatical constructions in writing.
2. Increased pupil confidence when writing in general.
3. Improved declarative knowledge of grammatical constructions.

Outcomes 1 and 3 were measured, however measuring pupil confidence (Outcome 2) was considered beyond the scope of the trial. Outcome 1 was measured via a writing composition assessment at the end of the trial, when teachers were intended to have completed ten weeks (20 sessions) of teaching using examples. Outcome 3 was measured via a Noun Phrase Grammar Assessment (NPGA) after the first five weeks of the trial, when teachers were intended to have completed the teaching related to noun phrases (ten sessions). As these teaching timelines were much more extended than is typical for research on worked examples (see 'Background' section above), it was not possible to predict whether these time frames are sufficient to see long-term outcomes.

Increased control and sophistication of use of grammatical constructions in writing was measured by a writing composition assessment, the bespoke Text-type Specific Writing Assessment (TSWA). This is a meaningful holistic outcome for Key Stage 3 teachers and pupils as it is a common type of task for measuring writing assessment across Key Stage 3 and Key Stage 4. The assessment measures three constructs, which are common in secondary writing assessments: 'Sentence structure and text organisation'; 'Punctuation'; and 'Composition and effect'. Pupils' *sophistication* of use of grammatical constructions is assessed within 'Sentence structure and text organisation'. As it is important that use of grammatical constructions is appropriate to the writing purpose and impacts the reader, we also assessed pupils' *control* of use of grammatical constructions through the 'Composition and effect' of the written text.

Improved declarative knowledge of grammatical constructions was measured by a closed-response grammar assessment, similar to the Key Stage 2 national curriculum assessment for English (e.g. the ability to identify an adjective, underline a noun phrase, or identify an error in an isolated example). Declarative knowledge is related to but not a prerequisite for ability to use grammatical constructions. Declarative knowledge may help pupils to consciously think about their writing and make deliberate choices, which may result in better writing, however, pupils can use implicit knowledge of grammatical constructions to replicate them, without any declarative knowledge of the grammatical construction (Watson, Newman, and Morgan, 2021). Therefore, declarative knowledge is treated as a possible secondary outcome of teaching with grammar patterns, as the teaching focuses on the use of grammatical constructions, rather than building declarative knowledge.

The primary impact research question is:

RQ1. What is the difference in writing composition[2] of Year 7 pupils taught using the three different approaches, as measured by a bespoke TSWA?

[2] Combined attainment in: i) 'Sentence structure and text organisation'; ii) 'Punctuation'; and iii) 'Composition and effect', across two text types.

We also estimate differences in writing composition for FSM-eligible pupils, as the logic model identifies that social disadvantage (i.e. FSM-eligibility) is systematically associated with lower prior attainment, and that use of worked examples is expected to more greatly impact pupils with lower prior attainment:

RQ1a. How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils?

A secondary impact research question addresses proximal impacts (at the end of the first block rather than the end of the trial[3]), and also estimates differences for FSM-eligible pupils in addition to those for all pupils:

RQ2. What is the difference in knowledge of grammatical constructs of Year 7 pupils taught using the three different approaches, as measured by an end-of-block NPGA?
RQ2a. How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils?

Further secondary research questions address heterogeneity in impacts by prior attainment and dosage. The first of these considers the theorised differential effect of the use of worked examples for pupils with varying levels of prior attainment. Although this may also be inferred by the findings for research questions 1a and 3 includes prior attainment specifically, rather than using FSM-eligibility as a proxy. We also considered this for a subgroup of EAL pupils, as the logic model identifies this group in particular as potentially benefiting from the use of worked examples (which is a common strategy in second language learning). However, this group is heterogeneous in terms of English language proficiency, with the least proficient pupils theorised to benefit most from the use of worked examples. Therefore, we will consider them as a subgroup in the context of this specific research question.

RQ3. How do the differences in Year 7 pupils writing composition vary by prior attainment (when measured by a bespoke TSWA)?
RQ3a. How do the differences in Year 7 EAL pupils writing composition vary by prior attainment (when measured by a bespoke TSWA)?

The final secondary research question is included to determine whether differences vary by the number of sessions taught and is included as a continuous measure of dosage under ITT analysis (separate compliance analysis to estimate the Complier Average Causal Effect [CACE] is considered in the 'Methods: Statistical analysis' section below). This research question is aligned with the logic model's contextual factor addressing teacher fidelity:

RQ4. How do the differences in Year 7 pupils writing composition vary by the number of sessions taught (when measured by a bespoke TSWA)?
RQ4a. How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils?

IPE

The IPE aimed to contextualise the impact findings, by exploring usual practice, and experiences and outcomes of the three approaches. It also aimed to explore teachers' experience of participating in a Teacher Choices trial, as this is a new and developing strand of work by the EEF.

RQ5. How, and how well, are the choices implemented? (Fidelity, adaptation, differentiation)
- Do teachers adhere to their allocated choices?
- Do teachers implement their assigned choice with fidelity?
- How does implementation vary (e.g. do teachers adapt the approaches to suit their context)?
- How different are these choices from teachers' usual practice?

RQ6. How well do teachers and pupils respond to the different choices? (Responsiveness)

[3] See 'Primary outcome' section below for further details regarding the rationale for choosing these two points at which to measure attainment.

- How do teachers respond to their allocated choices?
- What is the perceived engagement of pupils across different choices?

RQ7. What are the perceived outcomes of the different choices? (Perceived impact, moderators)

- What are the perceived outcomes for pupils (e.g. sophistication of grammar choices, confidence)?
- What are the perceived outcomes for teachers?
- Do perceived outcomes differ for specific groups of pupils (e.g. FSM, lower attainers)?

RQ8. To what extent did the trial design enable teachers to enact their allocated choices? (Time costs, mediators)

- To what extent does the teacher guidance/materials enable teachers to use their allocated choice?
- What guidance do teachers perceive they need (a) to use the choices within the trial, and (b) to continue using the choices beyond the trial?
- Were there any challenges in implementing different choices within the same school (e.g. contamination)?
- How does participating in the Teacher Choices trial affect teacher workload (e.g. planning, changing pedagogy)?
- What approaches did teachers use in the rest of their teaching time (e.g. compensation by use of different choices)?

Ethics and trial registration

The trial has been designed, conducted, and reported to CONSORT (Consolidated Standards of Reporting Trials) standards (http://www.consort-statement.org/) and in accordance with NFER's Code of Practice. All of NFER's projects abide by its Code of Practice, which is in line with the Codes of Practice from BERA (British Educational Research Association), MRA (Market Research Association), and SRA (the Social Research Association), among others. NFER is committed to the highest ethical standards in all of its activities and ethical considerations are embedded in its detailed quality assurance processes. In addition, ethical approval was obtained through the University College London (UCL) research ethics process in December 2023 (reference REC1914_AT) due to partnering with UCL, Birkbeck College London, and the University of Exeter.

Each participating school's headteacher provided their agreement to participate in the trial by signing the Memorandum of Understanding (MoU) that outlines the responsibilities of all parties involved in the trial. The head of English or other appropriate staff member acted as a key contact person for the trial. NFER shared a letter for Year 7 English teachers with full details about the trial, including the opportunity to withdraw from the trial. NFER also shared a parent letter and withdrawal form with schools to be sent to parents/carers of all Year 7 pupils. Through the withdrawal form, parents/carers had the opportunity to withdraw their child from the evaluation and associated data processing at any stage of the trial. A total of 13 pupils were withdrawn by their parents during the trial. A separate agreement process was used for the pupil focus groups and applied only to those selected to participate. This included an opt-out letter sent to schools for parents/carers of selected pupils, and an opt-in pupil agreement form collected before the focus group on the day of the school visit.

Any assessment where pupils construct text creates the potential for pupils to make safeguarding disclosures. We contacted all schools during the Summer Term to obtain contact details in advance for their designated safeguarding lead, and a deputy contact, as the marking period coincided with school summer holidays. During the training for our specialist writing markers, we asked markers to refer any scripts, which could potentially raise a safeguarding concern, so that these scripts could be considered by the project safeguarding team. Overall, 37 scripts were referred by markers for review by the team. This team included experienced English teachers, familiar with marking creative writing, and NFER's designated safeguarding lead, who met daily to read through these scripts and determine if they needed to be referred to the school. In total, scripts for less than ten pupils were referred to their school due to safeguarding concerns. In each case, the school confirmed in writing that they received the information they needed and would take any necessary action. Scripts from a further small number of pupils (less than ten), although not safeguarding issues specifically, were flagged with the relevant schools for content using a similar approach.

The trial was registered at the International Standard Randomised Controlled Trial Number (ISRCTN) registry, and given the registration number 14181429.

Data protection

All data gathered during the trial has been held in accordance with the data protection framework created by the Data Protection Act 2018 and the General Data Protection Regulation (GDPR) 2016/679 (GDPR, 2016) and treated in the strictest confidence by NFER, University of Exeter, Birkbeck, UCL, and the EEF. No individual or schools are identified in the report.

NFER were data controllers for the duration of this trial.

The legal basis for processing personal data was covered by: GDPR Article 6 (1) (f) which states that 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data' (GDPR, 2016). We carried out a legitimate interest assessment, which demonstrates that the evaluation fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities) and it has broader societal benefits. Therefore, it is in our legitimate interest to process and analyse personal data for the administration of this randomised controlled trial.

Privacy Notices were available for participating schools and teachers, and parents/carers of participating pupils.

As part of the sign-up process, NFER collected the name, contact details, and the job role of headteachers and the nominated trial lead.

NFER collected the following data for teachers participating in the trial: name; contact details; job role; gender; teaching qualifications and experience; current classroom practice and confidence in teaching; views of the grammar approach; trial materials; trial; and a use log for their allocated grammar approach. The University of Exeter asked a small number of teachers to review the trial materials in advance, collecting information about their views on these.

NFER collected personal data about pupils from participating schools. This included name, date of birth, unique pupil number (UPN), information on attainment grouping, responses from the grammar assessment (NPGA) and writing assessment (TSWA), and (for the sample of pupils in group discussions) attitudes towards the grammar approaches.

NFER also matched the pupils to the Department for Education's (DfE's) NPD to gain their gender, Key Stage 2 English attainment data, FSM-eligibility, and EAL status, to inform the analysis.

Within three months of the end of the project, NFER will send school and pupil data to the EEF's data archive partner. At this point, the EEF's data archive partner will keep a copy of the data and the EEF will become the data controller. Please see the EEF's archive guidance **here** and its Privacy Notice **here** for more information on how the EEF processes and will use personal data. NFER will retain personal data for one year after report publication in case there are any queries about the report. One year after the report publication, all personal data will be securely deleted.

Project team

Table 5 provides a list of the members in the project team, their institution, and their role in the trial.

Table 5: List of members in the project team and their institution and role

Name	Institution	Role
Helen Poet	NFER	Trial director
Katherine Aston	NFER	Trial manager
Andrew Smith	NFER	Impact design/analyst
Jose Liht	NFER	Trial statistician
Kathryn Hurd	NFER	Operations lead
Katharine Stoodley	NFER	Operations researcher
Lydia Wallis	NFER	Operations researcher
Sarah Millar	NFER	Operations researcher
Holly Critchley	NFER	Operations researcher
Eleanor Bradley	NFER	IPE researcher
Frances Brill	NFER	Outcome design
Rob Ager	NFER	Outcome design and marking
Sarah Gibb	NFER	Outcome marking
Katharine Larkin	NFER	Outcome marking
Vrinder Atwal	NFER	Project coordinator
Andrew Tolmie	UCL	Cognitive science expertise
Michael Thomas	Birkbeck	Cognitive Science expertise
Annabel Watson	University of Exeter	Grammar and English teaching expertise, teacher guidance lead

Study Advisory Board

Table 6 provides the members of the Study Advisory Board and their institution.

Table 6: List of members in the Study Advisory Board and their institution

Name	Institution
Bob Pritchard	EEF Dissemination and Impact Team
Gaia Sceriff	Oxford University
Joshua Clarke	Furze Platt Senior School
Lisa-Maria Müller	Chartered College of Teaching
Niki Kaiser	Norwich Research School
Steve Higgins	Durham University

Methods

Trial design

Table 7: Trial design

Trial design, including number of arms		Cluster randomised controlled trial (three arms)
Unit of randomisation		Teacher-class unit
Stratification variables (if applicable)		School
Primary outcome	Variable	Writing composition (combined attainment in: i) 'Sentence structure and text organisation'; ii) 'Punctuation'; and iii) 'Composition and effect', across two text types)
	Measure (instrument, scale, source)	Text-Type-Specific Writing Assessment (TSWA), developed by NFER, administered at the end of the trial period
Secondary outcome(s)	Variable(s)	Knowledge of grammatical constructs (focusing primarily on the noun phrase and its constituent parts)
	Measure(s) (instrument, scale, source)	Noun Phrase Grammar Assessment (NPGA), developed by NFER, administered at the end of the first block
Baseline for primary outcome[a]	Variable	Grammar, punctuation, and spelling attainment at Key Stage 2
	Measure (instrument, scale, source)	Marks achieved in grammar, punctuation and spelling at Key Stage 2 (KS2_GPSPAPER1MRK)
Baseline for secondary outcome[a]	Variable	Grammar, punctuation and spelling attainment at Key Stage 2
	Measure (instrument, scale, source)	Marks achieved in grammar, punctuation and spelling at Key Stage 2 (KS2_GPSPAPER1MRK)

[a]*See 'Outcome measures' section for more information about the choice of a baseline measure.*

The study used a cluster randomised design with the randomisation of Year 7 English teacher-class units to one of three trial arms:

1. Systematic use of worked examples (systematic worked arm).
2. Responsive use of worked examples (responsive worked arm).
3. Use of non-worked examples (non-worked arm).

Teacher-class units comprised a teacher and all participating Year 7 English classes taught by them. Where teachers shared teaching of classes, then multiple teachers were included in the same teacher-class unit (to prevent a class being taught by two teachers allocated to different trial arms). A teacher-class unit randomised (rather than school randomised) design was chosen as this required a smaller number of schools for a given minimum detectable effect size (MDES), thus, aiding recruitment. It also meant any given school would have teacher-class units randomised to more than one approach. Feedback from teachers during the scoping phase focus groups suggested that teachers would prefer this so that their school could try out and compare different approaches, rather than all teachers being allocated to the same approach, such as in a school randomised design. While the teacher-class unit randomised design was therefore, appropriate to the evaluation context, randomising teachers or classes within the same school may be associated with an increased risk of contamination (e.g. teachers sharing practices from their trial arm with each other) compared with a school randomised design. To minimise contamination in this context, individual teachers only received the guidance for their allocated approach and were asked to avoid discussing their approach with teachers allocated to a different approach, until after the trial teaching was complete. The IPE findings did not suggest that contamination was a substantial issue: 78% of teachers said that they had followed their allocated approach in their teaching sessions (endpoint survey); and all case study schools (n=6) indicated they had avoided discussing their allocated teaching approaches.

The trial's primary outcome was an end-of-trial measure of writing composition (i.e. TSWA), which included assessment of grammatical proficiency. This used a teacher-administered two-part pen-and-paper writing composition task. The secondary outcome was a closed-response measure of attainment in grammar (i.e. NPGA, focusing primarily on the noun phrase and its constituent parts) administered after the first five-week teaching block, which focused on noun phrases in narrative fiction. Both assessments were designed by the evaluation team as no suitable alternative assessment of Year 7 grammar could be identified. Full details, along with a discussion of the baseline measure, can be found in the 'Outcome measures' section below.

Participant selection

Eligible participants were teachers of Year 7 English and their pupils in state-funded secondary schools. All state-funded secondary schools in England were eligible to participate in the trial. During recruitment, we communicated that we would ideally like all Year 7 English teachers and classes in each school to participate (including where a teacher teaches more than one Year 7 English class). This was in order to maximise the statistical power of the trial given the number of schools participating. However, we appreciated the many burdens on schools and have therefore, accepted schools entering fewer than their total number of Year 7 English teachers and classes (thus, there was no restriction on school eligibility depending on the number of participating teachers in a school). The following selection criteria were used for teacher and class participation in this trial:

- teachers may teach one or more Year 7 English classes; but
- classes need to have either: i) one main English teacher; or ii) two main English teachers. In the latter case, both teachers are assigned the same approach to teaching grammar.

NFER was responsible for recruiting schools for the trial, with a recruitment target of 65 schools. We estimated this would yield 390 teachers (in teacher-class units, number unknown at the recruitment stage) eligible for randomisation (and approx. 11,440 pupils). Full details of the size of the sample for recruitment can be found in the 'Sample size' section below, along with details of the sample randomised and analysed.

During the 2023 Autumn Term, NFER publicised the trial to schools through newsletters and social media channels. During the 2024 Spring Term, NFER directly contacted a sample of state-funded secondary schools to invite them to participate in the trial. The EEF supported recruitment efforts by promoting the trial through their newsletters and social media channels. Interested schools completed an online Expression of Interest to help NFER to ascertain their eligibility for the trial. Eligible schools were then sent the school information sheet and an MoU. Schools signed up to the trial by the headteacher signing the MoU and providing the name of a key project contact (typically the head of English) to act as the coordinator of the trial in the school. At the end of the trial, schools received a payment of £100 per participating teacher or class (whichever number was higher) as a 'Thank You' for planning for their allocated teaching approach and organising the pupil evaluation activities.

Outcome measures

Baseline measures

In order to reduce schools' data collection burden and to avoid the need to develop a bespoke baseline assessment, we used administrative data from Key Stage 2 statutory assessments as baseline measures of attainment. As we designed bespoke measures of the primary and secondary outcomes, we did not have any pre-existing data to help us understand the likely correlation between our bespoke assessments and the Key Stage 2 assessments. We therefore, determined which of two available baseline measures better correlated with the outcome measure during preliminary analysis (see 'Outcomes and analysis' section 'Correlation between prior attainment and the TSWA and NPGA' subsection below), prior to entering treatment status into our analytical model. Our project study plan stipulated that the measure with the highest correlation would be chosen for subsequent analysis. We chose from the following two measures:

1. Marks achieved in the GPS test, which assesses grammar, vocabulary, and punctuation (KS2_GPSPAPER1MRK).
2. Marks achieved in English reading test (KS2_READMRK).

Ideally, we would have also been able to include a suitable Key Stage 2 writing assessment in this choice, but the teacher assessed component of the Key Stage 2 writing national assessment does not offer the required granularity to effectively correlate with our outcomes, given that it results in a reported categorical outcome (in the NPD; i.e. EXS, Working at the expected standard; WTS, Working towards the expected standard; and GDS, Working at greater depth within the expected standard) rather than marks or a scaled score.

Primary outcome

The study's **primary outcome measure** was the **TSWA**, see Appendix C. A bespoke outcome measure was necessary because there were no validated curriculum-aligned measures of the two-specific writing genres covered in the trial (narrative description and persuasive speeches). The TSWA is a two-part pen-and-paper writing composition assessment administered by teachers to Year 7 pupils at the end of the trial. The end of the trial was chosen as the appropriate time point because it meant that the pupils were undertaking the assessment after the period of ten weeks in which they were to have experienced one of the three teaching approaches and *both* blocks of grammar teaching (i.e. the 'noun phrases' grammar constructions in the first half-term block, and the 'clause/sentence' constructions in the second half-term block). The TSWA comprises two short writing tasks, with 20 minutes given for pupils to complete each task (approximately 45 minutes in total, including task introduction time). The pupil scripts were marked using the TSWA mark scheme by subject-specialist markers recruited by NFER. To assure the quality of marking, markers were trained and standardised in-house before completing any marking and supervised by expert marker managers.

The TSWA was developed by the NFER English assessment team before the start of the trial (October 2023 to March 2024). It was constructed from publicly available national curriculum assessment materials ('past papers') that were originally used in the Key Stage 2 (Year 6) national writing test in England between 2003 and 2012.[4] The two-specific short tasks were selected to align closely with each of the two writing genres—narrative fiction and persuasive speeches—underpinning the trial's two blocks of grammar teaching. In their original usage, these short tasks were each paired with an accompanying longer task. For adapted use in this trial with Year 7 pupils, the two shorter tasks were paired together.

The first task ('It's a Mystery') invited pupils to describe the opening scene of a mystery story. The task was designed to elicit the production of rich, descriptive language, including expanded noun phrases, within a fictional narrative frame. As such, the task corresponded with the teaching focus on the noun phrase grammar constructions in the first five weeks of the trial (i.e. the first half-term block on **narrative fiction**).

The second task ('Charity Choice') asked pupils to write a short speech to persuade their class to support a particular charity. It was designed to allow pupils to demonstrate their abilities to produce language that supports argumentation, including grammatically complex sentences and 'short sentences' for effect. This task, therefore, fitted well with the teaching focus on clause and sentence-level constructions in the second five weeks of the trial (i.e. the second half-term block on **persuasive speeches**).

The writing prompts themselves were used in the trial without adaptation. As part of their original development as national assessment instruments, these tasks were extensively and rigorously trialled in terms of their psychometric properties with large, national samples of Year 6 pupils, reviewed by consultative panels and adjusted for suitability prior to public use. During the development process for the writing prompts' use in this trial with Year 7 pupils, a small amount of informal trialling took place in one school, with a small group of Year 7 pupils. Each pupil completed the writing assessment and provided verbal feedback to an NFER researcher from the English assessment team. Pupils were from a mixed-ability class and were selected by the teacher to provide a spread of attainment. In addition, statistical tests of internal reliability were run on the study TSWA data (see 'Impact evaluation results: Statistical analysis: psychometric analysis' section below).

[4] See: http://www.satspapers.org/englishKS2SATS.htm

For optimum alignment, we used adapted versions of the original mark schemes. In the original versions, the first strand assesses the writing for 'Sentence structure, punctuation, and text organisation', using a two-bullet descriptor that requires the marker to make a judgement balancing the pupil's performance of 'Sentence structure and text organisation' with 'Punctuation'. Given that the best-fit judgement may mask pupils' grammatical performance (i.e. in cases where a pupil's punctuation performance on a given task is notably lower than their grammatical performance), we 'split out' the punctuation bullet into a separate strand for the purposes of the trial. This means that the adapted mark scheme used in the trial has three strands rather than two: i) 'Sentence structure and text organisation'; ii) 'Punctuation'; and iii) 'Composition and effect'.

We made small adjustments to the mark structure for all strands to allow markers to better reflect pupil performance at the top end of the ability range, given that Year 7 pupils may score more highly than Year 6 pupils in writing composition. Specifically, for 'Sentence structure and text organisation' and 'Punctuation', we introduced an extra mark, which represents performance above the descriptor in Band 4. For 'Composition and effect', we introduced one extra mark within Band 5 (to enable markers to distinguish between lower and higher performance in that band) and an extra mark above Band 5, which represents performance above the descriptor in Band 5.

Table 8 and Table 9 show the overall assessment structure and mark structure for the TSWA.

Table 8: TSWA overall assessment and mark structures

Writing task	Assessment strands	Marks
1. It's a Mystery (description in narrative)	Sentence structure and text organisation	0–5
	Punctuation	0–5
	Composition and effect	0–10
	Total available marks Task 1	*20*
2. Charity Choice (persuasive speech)	Sentence structure and text organisation	0–5
	Punctuation	0–5
	Composition and effect	0–10
	Total available marks Task 2	*20*
	Total available marks for TSWA	**40**

Table 9: Mark structure for the TSWA

Sentence structure and text organisation	Marks	Punctuation	Marks	Composition and effect	Marks
Band 1	1	Band 1	1	Band 1	1
Band 2	2	Band 2	2	Band 2	2–3
Band 3	3	Band 3	3	Band 3	4–5
Band 4	4	Band 4	4	Band 4	6–7
Performance above Band 4	5	Performance above Band 4	5	Band 5	8–9
				Performance above Band 5	10

Using this assessment structure allowed measurement and analysis of pupils' task performance in terms of different aspects of their writing abilities (i.e. 'Sentence structure and text organisation', 'Punctuation', and 'Composition and effect') in addition to a total score reflecting their writing performance more holistically. It also allowed for separate measurement of their performance on each text-type specific element (i.e. the description in the narrative task and the persuasive speech

task). Thus, the assessment could provide the basis for understanding the 'repertoire' of grammatical choices that the pupils used in creating their responses to the task prompts, both in terms of the relative sophistication of the constructions they used and the text-type appropriacy of the grammatical selections they have made. These were, however, not explicit foci of our analysis for this study. Before proceeding with impact analyses, we conducted psychometric testing of the functioning of the assessment tasks and items (for both primary and secondary outcomes; see 'Statistical analysis' section below).

Secondary outcomes

The study's **secondary outcome measure** was a pen-and-paper, closed-response assessment of grammar, focusing primarily on the noun phrase and its constituent parts: the **NPGA** (see Appendix D). A bespoke measure was needed to ensure the assessment content matched the grammar teaching content (noun phrases) in the first teaching block. It was administered by teachers at the end of the first five weeks of the trial as a proximal outcome (both temporally and in terms of content). This was chosen as the appropriate time point rather than at the end of the ten-week period so that the pupils would undertake the NPGA assessment directly after the five-week block where the teaching focus was on the noun phrase. This allowed the focus at the end of the ten-week period to be on the primary outcome measure. It guarded against any possible contamination from the NPGA that might have occurred were the NPGA to be conducted at the end of the ten-week period (e.g. the inadvertent priming of pupils to include expanded noun phrases in their writing compositions). This assessment contained up to 30 grammar items, and pupils were given approximately 30 minutes to complete the assessment.

It is important to recognise that this assessment was *not* intended to be a proxy measure of pupils' writing composition abilities. Rather, it was intended to provide some insight into pupils' abilities to *recognise* explicitly some of the key features of grammatical structures that are found in narrative fiction, including the expanded noun phrase constructions and patterns that formed the basis of the first five weeks of the trial.

As no suitable measure was available, it was necessary to create an assessment for the purpose of the trial by adapting existing materials. The NPGA was constructed by the NFER English assessment team before the start of the trial (October 2023 to March 2024), largely from a selection of existing grammar assessment items originally used in the Key Stage 2 national curriculum English GPS test. These are from publicly available 'past papers'.[5] As with the TSWA above, the rationale for using selected individual assessment items from this source was that they have already been extensively and rigorously trialled in terms of their psychometric properties with large national samples of Year 6 pupils, as part of their original national curriculum assessment development.

In order to compile this assessment, we assembled a selection of items that focus on the noun phrase. While the vast majority of items were from the Key Stage 2 national curriculum grammar tests, it was necessary to supplement the selection, for reasons of coverage and/or level of demand. Some of these items (4 items) were from Key Stage 2 national curriculum Level 6 tests (i.e. designed for more able Year 6 pupils); some (2 items) from the Key Stage 1 national curriculum tests, and one was adapted by NFER from an existing item. Although the majority of the items in the measure had been used previously, as noted above, it is important to bear in mind that the items had not been combined in this way before. Therefore, development included a small amount of informal trialling in the same school as the writing assessment trial, with a different small group of Year 7 pupils. Each pupil completed the grammar assessment and provided verbal feedback to a researcher from NFER's English assessment team. Pupils were from a mixed-ability class and were selected by the teacher to provide a spread of attainment. In addition, statistical tests of internal reliability were run on the study NPGA data (see 'Impact evaluation results: Statistical analysis: psychometric analysis' section below).

Sample size

To determine the required sample size for a three-arm cluster randomised controlled trial at the teacher-class unit level, we used a simulation approach, and a script developed specifically for this evaluation (written in R). We modelled a range of

[5] See http://www.satspapers.org/ks2english2016onwards.htm

scenarios, which considered varying assumptions (e.g. pre- and post-test correlations, number of teachers per school), and simulated 1,000 trials per scenario. Table 12 presents the assumptions, which we ultimately used to determine the sample size (design stage), along with the realised sample at randomisation and analysis.

Our calculations at the design stage were based on the equal allocation of teachers, classes, and pupils to the three trial arms, with all pupils being tested (for both primary and secondary outcomes). Based on our analysis of school and class sizes from publicly available data, we assumed six teachers of Year 7 English per school, and eight classes per school, with some teachers teaching more than one class. These assumptions were upheld by the data, which we collected from schools at the Expression of Interest stage of recruitment. However, as outlined in the 'Participant selection' section above, a school's eligibility for the trial was not contingent on the participation of six teachers and eight classes; in order to reduce the evaluation burden on schools our recruitment strategy allowed them to determine the number of teachers and classes taking part in the trial (although our communications encouraged schools to include all Year 7 teachers and classes, for the purpose of maximising statistical power). Based on publicly available data, we also assumed a class size of 22 pupils per Year 7 English class, 24.6% of whom are FSM-eligible (EVERFSM_6_P).

Our assumption about the pre- and post-test correlations was informed by the Key Stage 2 GCSE correlations reported by Singh *et al.* (2023). We further assumed that the Key Stage 2 Year 7 association would be more highly correlated due to the smaller time gap, although, as we were developing bespoke assessments for this evaluation, we did not have any pre-existing data to provide more accurate information (e.g. correlation will be contingent on the reliability of the Year 7 attainment measure). We therefore, chose an assumed correlation, which we believed to be realistic and slightly conservative (i.e. lower than might be realised in our analysis). As described in the 'Outcome measures' section above, we chose the most appropriate pre-test measure from the data available in the NPD to provide the best correlation.

We also assumed a nested variability structure of pupils within teachers/classes within schools. We selected a school-level ICC of 0.23, which is higher than the median (0.10) found by Singh *et al.* (2023) (although they found a maximum of 0.39). Our choice was informed by Demack (2019). Given the extent of streaming/setting we expected among our recruited classes, we assumed a relatively high class level ICC (0.40), also being guided by Demack (2019), which notes the magnitude of this parameter on account of the prevalence of streaming and setting practices in Key Stage 3 and Key Stage 4 English.

Our design stage estimates assumed 10% school-level attrition between recruitment and analysis, and of the remaining schools, 15% pupil-level attrition.[6] These estimates were based on our experience of recruiting to similar trials.

Randomisation

This study was designed as a three-arm randomised controlled trial, with teacher-class units randomly assigned to one of three treatment arms with equal allocations (1:1:1). Randomisation was stratified by school to balance the allocation to treatment arms within schools. Teachers were asked to use their allocated approach with their participating classes over a period of approximately ten weeks, for an average of twice a week, for approximately 15 minutes.

The decision to randomise at the teacher-class unit level was guided by sample size calculations, school recruitment considerations, and in discussion with teachers in previously held scoping phase focus groups (May 2023). For example, school-level randomisation would have required a larger number of schools (over 120) to take part in the trial, and teachers were concerned about all teachers within a school being allocated to the same trial arm as this would have meant that a school did not get the opportunity to participate in all three arms of the trial. Teachers also noted the difficulty associated with class-level randomisation, such that if different classes of pupils in a school are allocated to different treatment arms, this could potentially require a teacher to use different approaches to teaching grammar if they teach more than one Year 7

[6] Estimates based on randomisation data assumed 15% pupil-level attrition (no school-level attrition). Teachers were able to withdraw from the trial independently of their school. This had a potential implication for pupil attrition, depending on the timing of the teacher withdrawal and also whether the withdrawing teachers' classes were also taught by other teachers. We asked withdrawing teachers to still administer outcome measures wherever possible.

English class. Therefore, the randomisation of teacher-class units was preferred. This had clear implications where classes were shared among teachers; randomly allocating one teacher to a trial arm also implied the same allocation for other teachers who were teaching the same class.

Randomisation was stratified by school in order to balance the number of treatment arms within schools wherever possible (although we did not exclude schools with fewer than three classes from the trial). Although the teacher-class units differed in size, we expected the resultant number of pupils to be balanced across treatment arms due to random allocation. We did not further stratify by attainment grouping, as we expected that streamed classes would be balanced across treatment arms by the randomisation. Furthermore, our sample size calculation was based on the assumption of a class-level ICC, which would account for setting.

All randomisation was undertaken by a statistician to whom the identity of the recruited schools and teacher-class units was unknown. Randomisation was carried out in R and used syntax in order to ensure transparency and replicability (see Appendix L). The randomisation process was quality assured by two members of NFER; an evaluator who was also a member of the project, and a director who was not.

Recruitment in the 2024 Spring Term provided data about the number of schools, teachers, and classes participating in the trial. We revised our MDES estimates accordingly ('Randomisation' columns in Table 12, continuing to make the assumption that 24.6% of pupils would be FSM-eligible as this information was not available about individual pupils at this stage of the study (i.e. prior to accessing NPD data in the Office for National Statistics [ONS] Secure Research Service [SRS]). Data at the point of randomisation gave an estimated MDES of 0.30 for all pupils and 0.31 for FSM-eligible pupils (compared with 0.20 and 0.22, respectively, using the design stage assumptions). These increases were principally due to two factors: i) fewer participating teachers (driven by fewer schools being recruited than anticipated); and ii) larger teacher-class units. The second factor was due to a greater prevalence of shared teaching than we had assumed at the design stage, which therefore, resulted in fewer units for randomisation. For example, in some schools, the allocation of teachers to classes meant that all Year 7 English classes formed a small number of teacher-class units (e.g. one or two), rather than the six teacher-class units per school, which we had previously assumed. Although 0.20 is a conventional threshold for determining a trial sample size, there is evidence to suggest (e.g. Wolf and Harbatkin, 2023) that measures developed by researchers may be associated with larger effect sizes than broader standardised measures (which are often used in trials powered for an MDES of 0.20). The primary measure in this trial was narrow, closely reflecting the content of the teacher choice under investigation, and proximal in terms of the point of measurement relative to the timing of choice implementation. These factors provided some logical support to the argument that a comparatively large effect size may be observed in this trial. There is also evidence from published research literature about the use of worked examples, which suggests effect sizes exceeding 0.30 may be estimated (e.g. 0.48 in the meta-analysis undertaken by Barbieri *et al.,* 2023). As a result, despite the decrease in statistical power between the design and randomisation stages, the trial was considered to be adequately powered to go ahead. The lower class-level ICC in the analysed sample (relative to our assumption) also subsequently increased statistical power between randomisation and analysis.

Statistical analysis

Psychometric analysis of bespoke outcome measures

Before proceeding with impact analyses, we conducted psychometric testing of the functioning of the assessment tasks and items. This included work to determine the internal consistency of the measures and their correlation with Key Stage 2 attainment. We also checked the score distributions to ensure model assumptions were not violated.

Balance at baseline

To check for imbalance across the three treatment arms at baseline, we cross-tabulated the following characteristics of the pupils in the sample:

- FSM-eligibility;
- EAL status;

- gender;
- Key Stage 2 reading outcome; and
- Key Stage 2 GPS outcome.

We expected this check to confirm the correct functioning of the randomisation, and to also be useful given the potential complexity of randomising teacher-class units (e.g. crossover of teachers across classes).

We also assessed imbalance at baseline in terms of pupils' attainment by comparing the difference in means across the three treatment arms of the two Key Stage 2 baseline scores (outlined above). We fitted a pair of 'nested' multi-level models (two levels: pupil; and teacher-class unit), examining the difference in –2log-likelihood between a model with and without the two group dummies via a chi-squared test. The models for the Key Stage 2 GPS outcome are outlined below:

$$\text{KS2\_GPSPAPER1MRK}_{ij} = \beta_0 + u_{0j} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

$$\text{KS2\_GPSPAPER1MRK}_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

Where $\text{KS2\_GPSPAPER1MRK}_{ij}$ is the Key Stage 2 GPS score for the pupil i taught by the teacher-class unit j, u_{0j} is the random intercept for teacher-class unit j, and $school_k$ is a variable denoting the randomisation stratum of each teacher-class unit. The coefficients β_1 and β_2 of dummy variables x_{1j} and x_{2j} in model 2 represent the difference between the mean of the outcome variable for groups x_{1j} and x_{2j} and the mean for the reference group (non-worked examples), with β_0 representing the mean of the outcome variable for the reference group.

Equivalent models, replacing Key Stage 2 GPS outcome with Key Stage 2 reading outcome, are similarly defined.

Primary analysis

All primary and secondary analysis was undertaken on an ITT basis. The primary analysis **(research question 1)** investigated the impact of the different teaching approaches on writing composition measured using the end-of-trial TSWA. We used a likelihood ratio test to assess the null hypothesis that there is no difference in the mean scores between the three groups and to look at the difference in –2log-likelihood between a model with and without the two group dummies via a chi-squared test. To control for prior attainment, each pupil's Key Stage 2 baseline measure (as outlined above) was included in the model as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. This likelihood ratio test compared a pair of 'nested' models[7] as follows:

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

Where $TSWA_{ij}$ is the TSWA score for the pupil i taught by the teacher-class unit j, u_{0j} is the random intercept for teacher-class unit j, $KS2_{ij}$ is the baseline Key Stage 2 score for the pupil i taught by the teacher-class unit j and $school_k$ is a variable denoting the randomisation stratum of each teacher-class unit. Additionally, the coefficients β_1 and β_2 of dummy variables x_{1j} and x_{2j} in Model 2 represents the differences between the means of the outcome variables for groups x_{1j} and x_{2j} and the mean for the reference group (with β_0 representing the mean of the outcome variable for the reference group).

[7] All analyses were as specified in the project study plan, with the exception of those including interaction terms: our study plan did not include the terms '$\beta_1 x_{1j} + \beta_2 x_{2j}$' (representing the treatment arms) in the first lines of the nested models with interaction terms. This was an error, which has been corrected here. Model notation pertaining to models used in all analyses has been changed in this report to more accurately communicate the models used.

Where the likelihood ratio test resulted in a rejection of the null hypothesis (i.e. there are differences between the mean scores between the three groups—in this primary and all other analyses), we ran post-hoc tests to determine, which of the three teaching approaches differed and to calculate each of the three pairwise mean differences. These are reported as effect sizes and 95% confidence intervals (CIs) as set out below.

Secondary analysis

Secondary analysis **(research question 2)** was based on the NPGA, administered at the end of the first block, to understand the impact on English grammar attainment. Otherwise, the secondary analysis replicated the primary analysis outlined above, with the likelihood ratio test comparing a pair of 'nested' multi-level models as follows:

$$NPGA_{ij} = \beta_0 + u_{0j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

$$NPGA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

Where $NPGA_{ij}$ is the NPGA score for the pupil i taught by the teacher-class unit j, u_{0j} is the random intercept for teacher-class unit j, $KS2_{ij}$ is the baseline Key Stage 2 score for the pupil i taught by the teacher-class unit j and $school_k$ is a variable denoting the randomisation stratum of each teacher-class unit. Additionally, the coefficients β_1 and β_2 of dummy variables x_{1j} and x_{2j} in Model 2 represents the difference between the mean of the outcome variables for groups x_{1j} and x_{2j} and the mean for the reference group, with β_0 representing the mean of the outcome variable for the reference group.

Analysis of **research questions 3 and 4** followed a similar approach, adding interaction terms for prior attainment and dosage, respectively. For example, the model for research question 3 was as follows, again comparing a pair of 'nested' multi-level models where $KS2_{ij}$ is the prior attainment of the pupil i (in the teacher-class unit *j)*:

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \epsilon_{ij}$$

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \beta_4 KS2 * x_{1j} + \beta_5 KS2 * x_{2j} + \epsilon_{ij}$$

As per our primary analysis, where the likelihood ratio test resulted in a rejection of the null hypothesis, we planned to make post-hoc comparisons between treatment arms, calculating effect sizes and 95% CIs as set out below.

Estimation of effect sizes

Teacher Choices trials tend to have outcomes, which are much more proximal (e.g. topic tests, time spent planning a lesson, and tests that concern a particular component of learning) than those used in programme evaluations. Comparing the effect sizes of impact estimates, which are based on such different outcomes may therefore, not be informative and could potentially be misleading; we therefore, do not present effect sizes in the 'Executive summary' section of this report and instead present adjusted means with their CIs as being more meaningful for teachers in the context of the evaluation of the teaching approaches.

We do however, understand that calculating and presenting effect sizes may be useful for future research in this area, and have therefore, done so (where the likelihood ratio test resulted in a rejection of the null hypothesis), presenting these in the 'Impact evaluation results' section below. We have also used effect sizes as the basis for our sample size calculations (in terms of the MDES).

As specified in the EEF 2022 statistical analysis guidelines (EEF, 2022), the results of the analyses of both the primary and secondary outcome measures are reported as Hedge's *g*. In each case, the three pairwise mean differences were extracted from the full model used in the likelihood ratio test and converted to an effect size according to the formula:

$$g = \frac{\bar{o}_a - \bar{o}_b}{s^*}$$

where the subscripts *a* and *b* represent each of the two groups within each pairwise comparison, the numerator for the effect size calculations is equivalent to each pairwise mean difference, and the denominator is the unconditional total variance calculated by running a multi-level model for each outcome measure without covariates.

CIs for each effect size were computed by multiplying the standard errors of each pairwise mean difference by the 2.5th percentile of a pupil's *t*-distribution with the number of degrees of freedom associated with the sample size. The CIs for the coefficient were converted to effect size CIs using the same formula as the effect sizes themselves.

For both primary and secondary outcomes, we also present mean scores and score distributions in addition to calculating overall effect sizes. For the primary outcome we report mean scores separately for the two sub-tasks within the writing assessment (which align with the noun phrase grammar constructions and sentence grammar constructions) and for each assessment strand.

Subgroup analyses

Our main subgroup analyses **(research questions 1a, 2a, and 4a)** investigated the impact of the allocated approach on the writing composition of children eligible for FSM, using the FSM indicator *EVERFSM6* from the NPD, matched for analysis within the SRS. We approached this analysis in two distinct ways. First, we ran separate primary and secondary outcome analyses on a subset of FSM-eligible pupils only. As in the primary analysis, a likelihood ratio test was used to establish whether there are differences in the mean scores between the three treatment arms for this subgroup, and as per our primary analysis, where the likelihood ratio test resulted in a rejection of the null hypothesis we planned to make post-hoc comparisons between treatment arms, calculating effect sizes and 95% CIs as set out below. Second, we ran a likelihood ratio test that adds interaction terms to the models (i.e. models that include both the subgroup indicator and the product of the subgroup indicator and each treatment arm dummy variable). Both approaches conform to the EEF 2022 statistical analysis guidelines (EEF, 2022). The likelihood ratio test for the subgroup analysis with interaction terms is given by:

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \beta_5 FSM_{ij} + \epsilon_{ij}$$

$$TSWA_{ij} = \beta_0 + u_{0j} + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 KS2_{ij} + \sum_{k=1}^{k} \beta_k \, school_k + \beta_4 FSM_{ji} + \beta_5 FSM * x_{1j} + \beta_6 FSM * x_{2j} + \epsilon_{ij}$$

With FSM_{ij} being a dichotomous variable for the FSM-eligibility status of pupil i (in teacher-class unit *j)*, and the remaining variables as described in the 'Primary analysis' subsection above. Where the null hypothesis was rejected by the likelihood ratio test, we used a series of post-hoc analyses to identify the significant differences in interaction between subgroup indicator and treatment arm, as well as calculating effect sizes and their corresponding 95% CIs.

The sample size of EAL pupils with Key Stage 2 data was sufficient for us to analyse outcomes for EAL pupils (for **research question 3a** only) using the NPD indicator *LanguageGroupMajor_[term][yy]*. Our approach used models as per those of research question 3 on a subset of EAL pupils.

We used the same simulation approach described above to also estimate the MDES for subgroups, and the extent to which subgroup analyses were underpowered compared to the full sample. In accordance with the EEF 2022 statistical analysis guidelines (EEF, 2022), underpowered subgroup analyses are reported as exploratory.

Additional analyses and robustness checks

In order to determine the sensitivity of our primary research question findings to our model specifications, we specified the following additional analyses as variations to the nested models indicated for the primary analysis:

1. Models that include any pupil-level covariates that we found to be imbalanced at baseline or as a result of attrition.

2. Models which:
    i. omit the 'Punctuation' component of the TSWA (assess 'Sentence structure and text organisation', and 'Composition and effect' only); and
    ii. use 'Punctuation' only as the outcome.

    Finding an effect for the first but not the second would support an attribution to the treatment.

As we did not find any pupil-level covariates to be imbalanced at baseline (see 'Balance at baseline' section above), we did not proceed with the first of these additional analyses.

Analysis in the presence of non-compliance

A key recommendation from our scoping phase was to collect dosage/compliance data in a 'light-touch' manner, to: i) minimise additional teacher workload; and ii) increase the probability that teachers would complete and return compliance data. We did not, therefore, specify collecting this information at the pupil level, but instead at the class level using a one-sheet proforma for teachers to complete weekly. For each session, teachers were asked to select whether they used their allocated approach and the date of the session (see Appendix O). Although randomisation was at the teacher-class unit level, we defined compliance at the class level in order to simplify recording and to clarify the link between a classes' exposure to the allocated approach and pupil outcomes. For example, a class taught by more than one teacher had one proforma, and both teachers were allocated to the same treatment arm. Where one teacher was teaching two classes they used the same approach with both classes, and each class had its own proforma.

In order to try to increase the probability of teachers returning this data, we also asked the nominated project lead within each school to encourage them to do so. Within the window for returning session logs (July 2024), teachers were sent up to three reminder emails, and schools were called twice. Even with these reminders, response rates for this data remained lower than we expected: dosage data were missing for 48.8% of pupils, with only 34.5% of schools and 57% of teacher-class units returning complete data. The percentage of missing dosage data was correlated with the number of pupils in the teacher-class unit ($r=0.21$, $p=0.002$, $n=228$), which is likely to mean that teachers with more classes were less likely to return complete dosage data. There were higher proportions of missing dosage data in free schools (74.9%) and community schools (73.5%), compared with foundation schools (53%), sponsor led academies (45%), converter academies (43%), and voluntary aided schools (41%). While we considered approaching schools for missing logs in September 2024, we expected that most teachers who had not submitted a log had not completed them during the trial weeks. Therefore, we were concerned about the quality of data if teachers tried to complete them retrospectively.

Our project study plan specified that we would use the dosage data to:

- calculate the average level of compliance across classes and analyse the distribution of the compliance measure using histograms and boxplots (overall and by block); and
- estimate the CACE.

Our definition of compliance for the CACE analysis was to be in terms of the number of sessions (maximum = 20) in which a class was taught using the allocated approach, to define a 'full compliance' threshold (allocated approach used in 18–20 sessions), and a 'partial compliance' threshold (allocated approached used in 10–17 sessions). These definitions would allow for a CACE estimate to be obtained using instrumental variable modelling, for all pupils and separately for the FSM-eligible subgroup. Estimating impacts using alternate thresholds was included to allow us to consider the effect of different levels of class-level compliance on pupil outcomes.

However, given the large proportion of data missing for the dosage variable, which would be used in compliance analysis, and the fact that there was some evidence from the IPE endpoint survey that teachers' self-reported adherence may not accurately reflect their implementation, we did not ultimately consider the proposed CACE analysis to be viable. Following consultation with a Study Advisory Board member with methodological expertise and discussion with the EEF, we decided not to proceed with this analysis. Our impact estimates were therefore, limited to analysis on an ITT basis, which may underestimate the effects of the allocated approaches by including all pupils as randomised, irrespective of the number of sessions delivered using the allocated approach.

Missing data analysis

Our initial sample size calculations at the design stage included the assumption of attrition (and therefore, missing outcome data) for 10% of schools and 15% of pupils randomised. We understood this to be a conservative estimate based on our experience across a number of trials, and our school engagement strategy prioritised the retention of schools and teachers in the study. Of the 55 schools included in the randomisation, five withdrew fully from the evaluation, and two partially withdrew (agreeing, however, to continue with data collection), representing a school-level full withdrawal rate of 9.1%. Schools withdrawing from delivery or evaluation were asked about the reason, of those who provided a response, the most common reason was staffing changes. Of the 8,901 pupils randomised, however, only 6,247 were included in the primary analysis; the level of attrition at the pupil level being 29.3%. This is partially explained by the withdrawal of schools (10.4%, 928 pupils), and also due to data missing from those schools remaining; 28.9% of teacher-class units randomised returned primary outcome data for fewer than 90% of their pupils. The two main reasons for data missing at the pupil or class level were pupils being absent on the test day (655 pupils), and classes not completing the assessment (196 pupils, representing eight classes in four schools). This may be explained by changes in secondary school timetabling in the final weeks of the Summer Term, for example, school trips, sports events, and enrichment days. To maximise the return of primary outcome data, we used a flexible three-week testing window (1–19 July 2024), encouraging teachers to choose a suitable assessment date for each class. NFER test administrators visited each school to deliver and collect the assessment materials, in order to reduce the administrative burden on schools and support teachers with any queries. We also extended the planned assessment window, allowing schools to return completed assessments until the end of the Summer Term.

The 'Impact evaluation results: Statistical analysis: Missing data analysis' section below further quantifies the level of missingness for the primary outcome and other variables. Following this analysis, we investigated patterns of missing data by means of a two-level (pupil and teacher) logistic model where the outcome was missingness on the writing assessment, with baseline Key Stage 2 attainment and the school randomisation indicator as covariates. Additional variables that may have been associated with missingness, but which were not included in the primary analysis (i.e. gender, FSM-eligibility, EAL status, treatment arm), were also included as covariates. Given the findings of these analyses (described in the 'Impact evaluation results: Statistical analysis: Missing data analysis' section), we assumed that data were Missing at Random (MAR) and conducted multiple imputation as the basis for an analysis of research question 1 with the missing data imputed. The imputation was done using multi-level predictive mean matching (with five plausible values derived for each case). The primary outcome model was run on the resulting five datasets as a sensitivity check for the primary outcome analysis.

IPE

Research questions

RQ5. How, and how well, are the choices implemented? (Fidelity, adaptation, differentiation)

- Do teachers adhere to their allocated choices?
- Do teachers implement their assigned choice with fidelity?
- How does implementation vary (e.g. do teachers adapt the approaches to suit their context)?
- How different are these choices from teachers' usual practice?

RQ6. How well do teachers and pupils respond to the different choices? (Responsiveness)

- How do teachers respond to their allocated choices?
- What is the perceived engagement of pupils across different choices?

RQ7. What are the perceived outcomes of the different choices? (Perceived impact, moderators)

- What are the perceived outcomes for pupils (e.g. sophistication of grammar choices, confidence)?
- What are the perceived outcomes for teachers?
- Do perceived outcomes differ for specific groups of pupils (e.g. FSM, lower attainers)?

RQ8. To what extent did the trial design enable teachers to enact their allocated choices? (Time costs, mediators)

- To what extent does the teacher guidance/materials enable teachers to use their allocated choice?

- What guidance do teachers perceive they need (a) to use the choices within the trial, and (b) to continue using the choices beyond the trial?
- Were there any challenges in implementing different choices within the same school (e.g. contamination)?
- How does participating in the Teacher Choices trial affect teacher workload (e.g. planning, changing pedagogy)?
- What approaches did teachers use in the rest of their teaching time (e.g. compensation by use of different choices)?

Research methods

The IPE uses a mixed-methods approach, as described below and summarised in Table 10. Data collection included light-touch data from teachers on dosage (i.e. teaching sessions and whether they used their allocated approach), a teacher survey, and a case study approach to gather qualitative data within a subsample of schools. The case study visits involved observations of the different teaching approaches in practice, as well as interviews with participating teachers and pupil focus groups. This enabled triangulation and contextualisation across data sources, specifically drawing together observed teaching, and teacher and pupil views.

At the design stage, we intended to visit nine schools, with the aim of observing two teaching approaches per school. However, as the case study visits were carried out towards the end of the trial period, we faced a highly constrained time frame for visits (within the second five-week teaching block, and before endpoint assessments), which overlapped with national exams for Key Stage 4 and Key Stage 5. Despite reminders, we also had a low response rate to our visit requests. This meant our achieved sample for visits was six schools, lower than the nine schools we aimed for.

The case study's achieved sample, represented a range of school characteristics. This included whether grammar is usually taught separately or integrated into lessons (from the teacher baseline survey, in two schools, all teachers integrated their grammar teaching, while in the other four schools, teachers varied in integrating grammar teaching, teaching grammar separately, or not explicitly teaching grammar). The sample also included a varied proportion of FSM-eligible pupils (varied from c. 15% – c. 30%).

Teacher surveys

We conducted short online surveys of all class teachers at baseline (28 February 2024 to 28 March 2024) and endpoint (9 July 2024 to 26 July 2024). The baseline survey was sent to teachers before randomisation, and teachers were asked to complete it as soon as possible. Due to the rapid timeline for trial recruitment and randomisation, and as we did not expect teacher baseline responses to be influenced by their allocation, the survey stayed open for responses until the end of Spring Term, which for a small number of teachers could have been after they were notified of their allocation (26 March 2024).

A unique link was sent via email to each teacher, to link their responses across the surveys and to their classes' data. All questions were closed, and surveys took around ten minutes (baseline) and 15 minutes (endpoint) to complete.

Baseline survey questions focused on:

- teacher characteristics, including experience, main teaching subject, and degree specialism;
- teacher self-reported confidence and efficacy for teaching grammar; and
- usual practice in teaching grammar and using examples.

Endpoint survey questions focused on:

- fidelity (the extent to which teachers had adhered to the implementation guidance);
- any adaptations the teachers had made;
- teacher responsiveness and engagement;
- teacher self-reported confidence and efficacy for teaching grammar;

- perceptions of any challenges in implementing the teaching approaches; and
- perceived impact on pupils.

The baseline survey response rate was 62% (n=202). The endpoint survey was sent to all teachers who were still participating in the trial at endpoint (n=270), the response rate was 53% (n=144). For both surveys, the response rate reflected the short time frame for collection in a one-term trial, and the prioritisation of data being collected at the same time point, which was needed for the primary impact analysis (pupil identifier data and endpoint testing data, respectively).

Observations

We used an observation schedule to describe key aspects of teaching, including session length, how the examples related to the overall lesson, how examples were analysed and/or broken down into steps, and what pupils were asked to do. The observation schedule captured the fidelity of implementation of the allocated approach (e.g. whether the teaching session was separate or embedded in the lesson, the use of steps for worked example approaches, and any pupil writing in the non-worked example approach), and therefore, observers were not blinded to the allocated approach. The researcher also observed teacher preparedness and responsiveness during the session.

In terms of the 'Reach' of our observations, while we were able to observe multiple approaches in two schools, in the other four schools, Year 7 timetabling constraints meant we could only observe one approach. Across the six schools visited, we were able to observe at least two examples of each approach (four systematic worked, three responsive worked, and two non-worked).

Teacher interviews

We used a semi-structured interview schedule to explore: what had gone well; any challenges faced; the extent to which they followed or adapted guidance; how well pupils responded; and the perceived impact on pupils.

In total, we conducted six group interviews and one individual interview across seven schools, comprising 18 teachers. We were able to interview the teacher(s) of observed class(es) in all six case study schools. We asked other teachers participating in the same choice to join a group discussion to increase the number of teachers interviewed, and in two schools, we also heard views from teachers with a different allocated approach. In addition, we interviewed a lead teacher who was unable to participate in a full case study visit.

Pupil focus groups

Focus groups explored pupils' perceptions and experience of the relevant choice, and their perceptions of impact. We ran one pupil focus group of about five to six pupils in each of the six case study schools, focusing on one choice per school, with pupils from an observed class. We asked teachers to select pupils across the class range of attainment and gender, and, where possible, to include at least one FSM-eligible pupil. Due to unexpected changes on visit days, we were only able to conduct focus groups with pupils from the worked example arms (four pupil groups for systematic worked examples, and two pupil groups for responsive worked examples). As visits were near the end of the Summer Term (June 2025), it was not possible to schedule follow-up visits to talk to pupils who were taught with non-worked examples.

Analysis

Our analytical focus was the implementation and perceived outcomes for each of the three teaching approaches, including similarities and differences.

Qualitative data – observations, interviews, and focus groups

Observation notes were coded using MAXQDA. Interview and focus group data were fully transcribed and then coded using MAXQDA. We initially coded all data deductively based on the IPE dimensions, coding these to multiple dimensions where relevant. We then undertook a thematic analysis (Braun and Clarke, 2006) for each IPE dimension by inductively analysing the relevant text. Data relating to perceived outcomes were coded inductively to ensure we captured participants'

perceptions accurately and compared with the logic model during reporting. Reporting of key findings for each IPE dimension focuses on similarities and differences across the three teaching approaches.

We used the same coding framework for the observation, interview, and focus group data. We compared views across participant groups (pupils and teachers) and data type (comparing teacher interviews with observed practice), both for each case study school and for the sample as a whole.

Quantitative data – surveys, attendance registers, and observations

Quantitative IPE data from the surveys, attendance registers, and observations were primarily analysed using descriptive statistics, including percentage responses to different response options, or means and standard deviations (SDs), depending on the question type.

Results are provided separately for each choice/allocated approach, where cell counts allow, to explore similarities and differences. We compared teacher confidence and efficacy at baseline and endpoint to explore any differences, including comparing changes for each choice.

Triangulation of qualitative and quantitative data

We designed the surveys, observation schedule, and interview/focus group topic guides concurrently so that relevant data was captured using the most appropriate method. To integrate data, we created an analysis framework outlining the links between the IPE dimensions, each survey question, and themes from the qualitative data. We triangulated findings between the different approaches and used the qualitative case study data to help interpret the quantitative survey data.

Table 10: IPE methods overview

IPE dimension	Research question addressed	Teacher log	Surveys	Teacher interviews	Teacher observations	Pupil focus groups	Sample size and sampling criteria	Data analysis methods
Fidelity, Adaptation	5	x	x	x	x		All teachers (survey/log) plus case studies	Descriptive statistics (log/surveys) and thematic analysis (case studies)
Responsiveness	6	x	x	x	x	x		
Perceived impact, Moderators	7		x	x		x		
Differentiation, Costs, Mediators	8		x		x			

Timeline

Table 11: Set-up and evaluation timeline

Dates	Activity	Staff responsible / leading
September 2023 – December 2023	Start-up meeting with the EEF Recruitment warm-up activities, including the launch of Expression of Interest Agree curriculum content Develop teacher guidance and teaching materials	Helen Poet Katharine Stoodley Dr Annabel Watson
October 2023 – March 2024	Develop trial study plan and statistical analysis protocol Develop bespoke outcome assessment	Andrew Smith, Gemma Schwendel, Katherine Aston Frances Brill
January 2024 – March 2024	Recruit schools to the trial Design IPE instruments Informally trial teacher guidance and teaching materials	Katharine Stoodley Katherine Aston Dr Annabel Watson
March 2024 – April 2024	Randomisation Administer guidance to schools	Andrew Smith Katharine Stoodley
April 2024 – June 2024	Apply for NPD data Trial period Tests administered by schools IPE activities	Andrew Smith Katharine Stoodley Katherine Aston, Eleanor Bradley
June 2024 – July 2024	Schools submit test data	Katharine Stoodley
August 2024	Data to DfE for NPD matching	Andrew Smith, Jose Liht
December 2024 – January 2024	Impact analysis IPE analysis	Andrew Smith, Jose Liht, Eleanor Bradley
February 2024 – April 2024	Draft report	Helen Poet, Katherine Aston
April 2024 – September 2025	Finalise report for submission	Helen Poet, Katherine Aston
October 2025	Submit data to the EEF data archive and update ISRCTN registry with results	Andrew Smith, Jose Liht

Impact evaluation results

Participant flow including losses and exclusions

The participant flow diagram for the trial is shown in Figure 2. At the recruitment stage, 55 schools with 8,930 pupils agreed to participate and provided pupil data. At this point, 29 pupils were lost to the sample due to a teacher withdrawing from the study, leaving 8,901 pupils to be randomised: i) 2,953 to the systematic worked examples (across 46 schools within 76 teacher-class units); ii) 3,253 to the responsive worked examples (across 50 schools within 76 teacher-class units); and iii) 2,695 pupils (across 51 schools within 76 teacher-class units) to the non-worked examples. These were then passed on to the SRS to be matched to their NPD records in order to add their Key Stage 2 prior attainment, FSM, and EAL status. After loss to non-response (no writing assessment completed), and inability to match 385 records to the NPD dataset, a total of 6,297 pupils across the three arms were included in the primary outcome analysis. Attrition consisted of 930 pupils from systematic worked examples, 837 from responsive worked examples, and 824 from non-worked examples.

Figure 2: Participant flow diagram

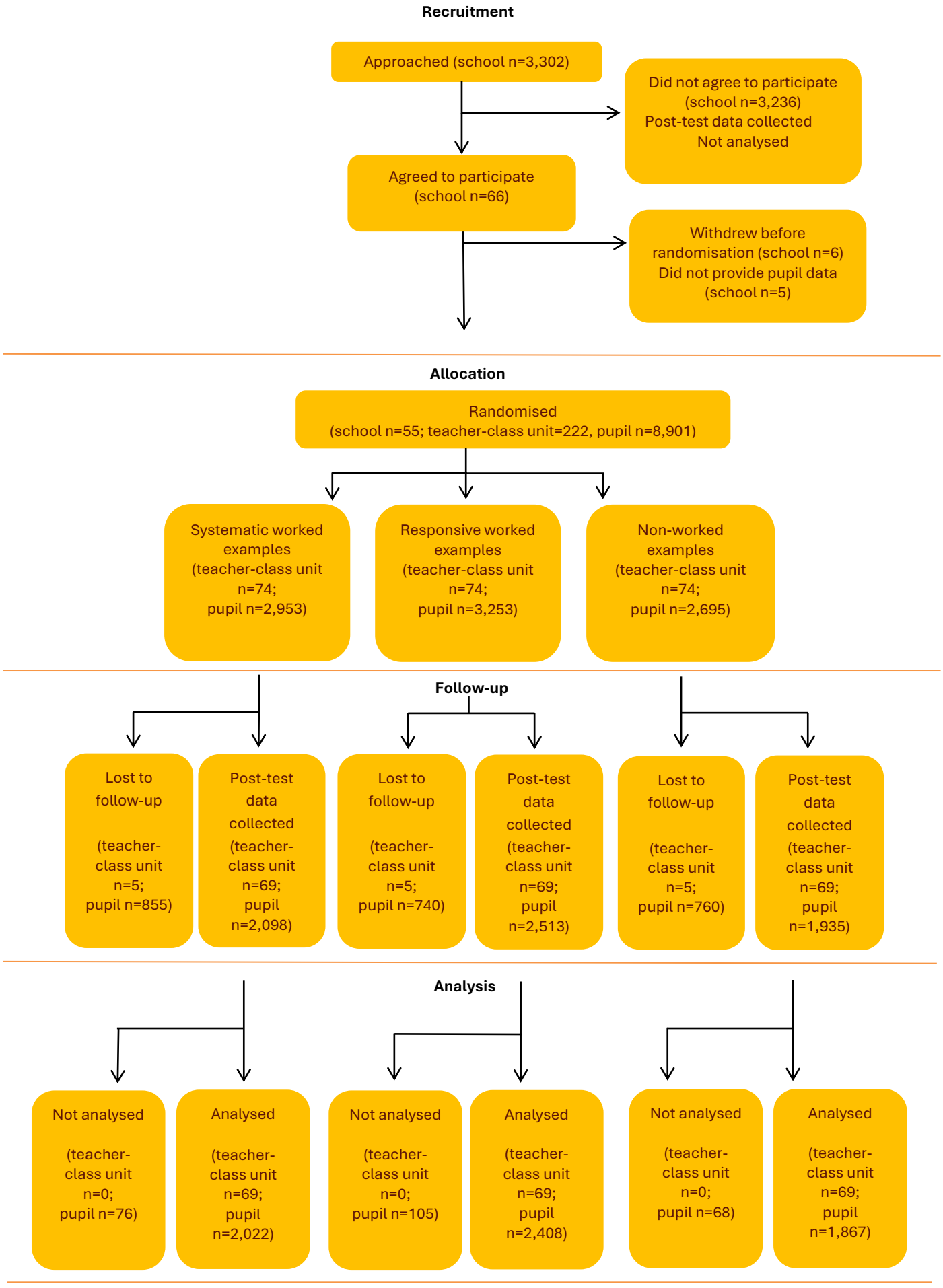


Table 12 shows the estimated MDES figures at the design, randomisation, and analysis stages. For the latter, the MDES estimate uses the achieved sample for the primary outcome (n=6,297) and subgroups. The ICC for the class level and the pre- and post-test correlations have been updated. The achieved sample was powered to detect effect sizes of 0.25 for all pupils, 0.28 for FSM-eligible pupils, and 0.30 for EAL pupils. Although the number of schools and pupils at the analysis stage was smaller than at the design stage, a noticeably lower ICC at the class level compensated for the reduction in sample size, making the anticipated power broadly similar in both cases.

Table 12: MDES estimates

		Design		Randomisation		Analysis		
		Overall	FSM	Overall	FSM	Overall	FSM	EAL
MDES		0.20	0.22	0.30	0.31	0.25	0.28	0.30
Minimum difference in means						1.49	1.67	1.90
Power		0.80	0.80	0.80	0.80	0.80	0.80	0.80
Pre-/ post-test correlations	Level 1 (pupil)	0.60	0.60	0.60	0.60	0.56	0.51	0.59
ICCs	Level 2 (class)	0.40	0.40	0.40	0.40	0.28	0.23	0.30
	Level 3 (school)[a]	0.23	0.23	0.23	0.23	–	–	–
Alpha[b]		0.05	0.05	0.05	0.05	0.05	0.05	0.05
One-sided or two-sided?		Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided	Two-sided
Average cluster (teacher-class unit) size		22	5.41	25.9	6.4	28.4	5.1	2.8
No. of trial arms		3	3	3	3	3	3	3
Schools	Systematic worked	65	65	55	55	41	40	37
	Responsive worked	65	65	55	55	44	42	38
	Non-worked	65	65	55	55	46	46	41
	Total	**65**	**65**	**55**	**55**	**50**	**50**	**50**
Teacher-class units	Systematic worked	–	–	74	74	69	65	56
	Responsive worked	–	–	74	74	69	67	59
	Non-worked	–	–	74	74	69	68	57
	Total	**–**	**–**	**222**	**222**	**207**	**200**	**172**
Pupils	Systematic worked	3,813	938	2,984	656	2,022	421	338
	Responsive worked	3,813	938	3,251	911	2,408	646	453
	Non-worked	3,813	938	2,695	737	1,867	473	330
	Total	**11,439**	**2,814**	**8,930**	**2,304**	**6,297**	**1,540**	**1,121**

[a]The simulations used to calculate the MDES at the design and randomisation stages used three-level models, which differed from the models used at the analysis stage (the latter were as per the project study plan).
[b]The alpha was set at 0.05 for the analysis of variance (ANOVA) F-statistic within the simulation syntax. The null hypothesis is that there is no difference between any of the means. Under this scenario, no adjustment for multiple comparisons is necessary as it is inherent in the F-statistic.

Attrition

The rate of pupil-level attrition for the primary outcome is reported in Table 13 below. As can be seen, the attrition ranged from 26% for the responsive worked arm to 31% for both the systematic worked and non-worked arms, with an overall attrition of 29%. In the overwhelming majority of cases, the reason for attrition was that the pupil did not complete the writing assessment (primary outcome measure), either because the school or class withdrew from the evaluation prior to testing, or the writing assessment was returned to NFER unused.

In a very small number of cases, parents withdrew their children after they had been randomised (13 pupils). A further loss of data resulted from the inability to match 385 pupils' records to the NPD dataset containing the KS2_ GPSPAPER1MRK prior attainment mark at the SRS/DfE merge stage, which was included in the primary outcome model. It thus, compounded the total number of missing cases at the analysis stage.

High attrition, and differential attrition across arms, can potentially introduce bias into analysis findings. To explore the impact of attrition on the findings, we ran missing data analysis and a sensitivity check using imputed data. The 'Statistical analysis: Missing data analysis' section below describes these analyses, and the robustness of the impact analysis against attrition threats.

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

	Allocation	Responsive worked	Systematic worked	Non-worked	Total
No. of pupils	Randomised	3,253	2,953	2,695	8,901
	Analysed	2,408	2,022	1,867	6,297
Pupil attrition (from randomisation to analysis)	Number	837	930	824	2,591
	Percentage	26	31	31	29
No. of teacher-class units	Randomised	74	74	74	222
	Analysed	69	69	69	207
Teacher-class unit attrition (from randomisation to analysis)	Number	5	5	5	15
	Percentage	7	7	7	7

Pupil characteristics

The pupil characteristics at the randomisation stage are presented in Table 14 below. The FSM-eligibility of pupils ranged from 22% for the systematic worked arm to 28% for the responsive worked arm and the non-worked arm, compared to the 2023 national average of 23%. The EAL status ranged from 15% for the systematic worked arm to 19% for the responsive worked arm, compared to the 2023 national average of 18%. In regard to gender, the sample ranged from 50% males for the systematic worked arm to 51% males for the responsive worked arm and the non-worked arm, compared to a national average of 54% males. The prior attainment as measured by the marks achieved in the English reading test KS2_READMRK (KS2 READING) ranged from a mean of 30.69 for the non-worked arm to 31.19 points for the systematic worked arm compared to a national average 32–33 points. The prior attainment as measured by the marks achieved in the KS2 GPS test KS2_GPSPAPER1MRK (Key Stage 2 GPS) ranged from a mean of 32.29 for the non-worked arm to 33.05 points for the responsive worked arm.

In the following section on arm balance, it can be appreciated that although there were important differences between the sample as randomised and the national averages, the randomisation was considered successful in minimising the differences in pupil characteristics across the study's arms at the analysis stage.

Table 14: Baseline characteristics of groups as randomised[a]

Pupil level (categorical)	National-level mean	Systematic worked		Responsive worked		Non-worked		
		n/N	Count (%)	n/N	Count (%)	n/N	Count (%)	
FSM – no	77	2,287	78	2,322	72	1,946	73	
FSM – yes	23	656	22	911	28	737	28	
(FSM – missing)		(9)		(12)		(8)		
EAL – no	82	2,485	85	2,616	81	2,244	84	
EAL – yes	18	448	15	603	19	428	16	
(EAL – missing)		(19)		(26)		(19)		
Gender – male	54	1,469	50	1,639	51	1,354	51	
Gender – female	46	1,474	50	1,594	49	1,329	50	
(Gender – missing)		(9)		(12)		(8)		
Pupil level (continuous)		n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	n/N (missing)	Mean (SD)	Effect size[b]
Key Stage 2 reading	32–33[c]	2,820 (132)	31.19 (10.21)	3,087 (158)	30.88 (10.02)	2,587 (104)	30.69 (9.93)	
Key Stage 2 GPS	Not available	2,826 (126)	32.83 (10.41)	3,088 (157)	33.05 (10.14)	2,589 (102)	32.29 (10.26)	

[a]*No school-level data is provided due to the need to adhere to Statistical Disclosure Control when reporting data processed in the ONS SRS.*
[b]*We have not included effect sizes as this would require multiple comparisons per row.*
[c]*National average reported as a range of raw values in:* ***www.gov.uk/government/publications/key-stage-2-tests-2024-scaled-scores/2024-key-stage-2-scaled-score-conversion-tables***

Statistical analysis

Balance at baseline

To assess imbalance at analysis stage between the three treatment arms at baseline we produced cross-tabulations of the background characteristics of the pupils in the sample and also ran nested multi-level comparisons of models for both prior attainment variables.

For the cross-tabulations, we used the following pupil-level characteristics:

- FSM-eligibility;
- EAL status;
- gender;
- Key Stage 2 reading outcome; and
- Key Stage 2 GPS outcome.

The tabulation for FSM-eligibility across the trial arms is shown in Table 14 above. As can be seen, the proportion of being FSM-eligible ranged between 28% for the responsive worked arm and 22% for the systematic worked arm. A test of independence between arm and FSM proportion showed no evidence of imbalance (p=0.918).

The tabulation for EAL across the trial arms is shown in Table 14 above. As can be seen, the proportion of being EAL ranged between 19% for the responsive worked arm and 15% for the systematic worked arm. A test of independence between arm and EAL proportion showed no evidence of imbalance (p=0.995).

The tabulation for gender across the trial arms is shown in Table 14 above. As can be seen, the proportion of being female ranged between 49% for the responsive worked arm and non-worked arm and 50% for the systematic worked arm. A test of independence between arm and gender showed no evidence of imbalance (p=0.999).

The means for Key Stage 2 GPS across the trial arms are shown in Table 14 above. As can be seen, the means ranged between 32.29 for the non-worked arm and 33.05 for the responsive worked arm.

The means for Key Stage 2 reading across the trial arms are shown in Table 14 above. As can be seen, the means ranged between 30.69 for the non-worked arm and 31.19 for the systematic worked arm.

The cross-tabulations for FSM-eligibility, EAL status, gender, Key Stage 2 GPS score, and Key Stage 2 reading score confirmed that the randomisation was successful in balancing out the background characteristics of the sampled pupils.

In addition to examining balance across the previous variables, we compared the mean size of the teacher-classroom units in a number of pupils across the treatment arms. Table 15 below shows that the mean size of the teacher-class unit ranged between 35.46 pupils for the non-worked arm to 42.80 pupils for the responsive worked arm. This imbalance, given the equal number of teacher-class units randomised to each arm and present at the analysis stage (after listwise deletion due to missing data), explains the difference in pupil numbers across arms (see Table 12, p. 38).

Table 15: Teacher-class unit size across arms

Arm	N	Mean	SD
Systematic worked	76	38.86	26.85
Responsive worked	76	42.80	28.38
Non-worked	76	35.46	27.06

As stated in the project study plan, we also assessed imbalance at baseline in terms of pupils' attainment by comparing the difference in means across the three treatment arms for the Key Stage 2 GPS score and Key Stage 2 reading score baseline measurements. In order to test the balance of Key Stage 2 GPS score, we used a nested multi-level model (two levels: pupil; and teacher-class unit) in order to examine the difference in –2log-likelihood between the model with and the model without the teaching approach dummies via a chi-squared test. The likelihood ratio test comparing the null model to the alternative model (including the teaching approach dummies) for the Key Stage 2 GPS score resulted in a chi-square of 2.43 with 2 degrees of freedom (p=0.296; n=8,446). This indicates that the alternative model did not provide evidence of a significantly better fit to the data compared to the null model, and thus, there is no evidence of imbalance in previous achievement across the teaching approaches as indicated by the Key Stage 2 GPS score variable.

Table 16: Key Stage 2 GPS score across arms

Arm	N	Mean (95% CI)	Standard error
Systematic worked	2,826	32.54 (30.93, 34.14)	0.81
Responsive worked	3,088	32.71 (31.15, 34.26)	0.79
Non-worked	2,589	31.38 (29.82, 32.94)	0.79

Furthermore, for the Key Stage 2 reading baseline score, a similar approach was taken. A multi-level model was fitted (two levels: pupil; and teacher-class unit) in order to examine the difference in –2log-likelihood between a model with and without the teaching approach dummies via a chi-squared test. The likelihood ratio test comparing the null model to the alternative model (including the teaching approach dummies) for the Key Stage 2 reading score resulted in a chi-square of 1.79 with 2 degrees of freedom (p=0.409; n=8,437). This indicates that the alternative model did not provide evidence of a significantly better fit to the data compared to the null model and thus, there is no evidence of imbalance in previous achievement across the teaching approaches as indicated by the Key Stage 2 reading score variable.

Table 17: Key Stage 2 reading score across arms

Arm	N	Mean (95% CI)	Standard error
Systematic worked	2,820	30.78 (29.17, 32.39)	0.82
Responsive worked	3,087	30.80 (29.24, 32.36)	0.79
Non-worked	2,587	29.72 (28.15, 31.28)	0.79

Both the cross-tabulations for the background variables as well as the nested multi-level comparisons of models for both prior attainment variables indicated that the treatment arms were successfully balanced by the randomisation and that there was no need to control for additional variables in the regression models when examining the research questions.

Psychometric analysis

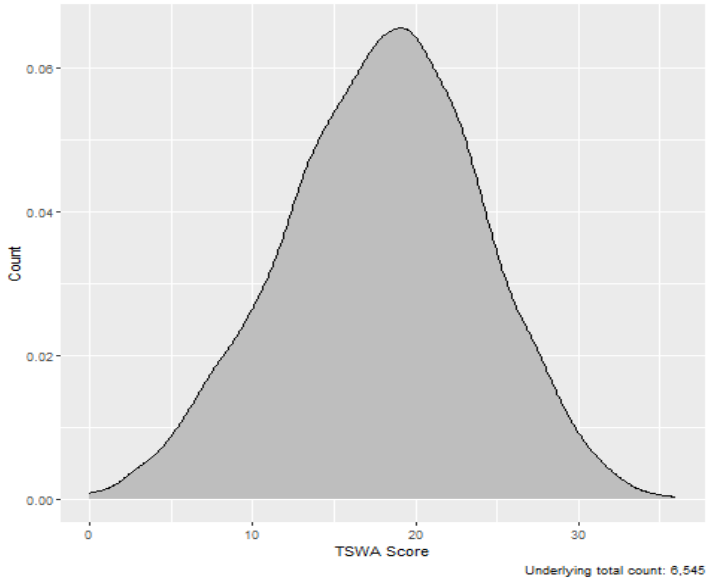
TSWA

The psychometric analysis of the writing assessment (TSWA) indicated an ICC of 0.73, which shows good intercoder agreement with an average item-total correlation of 0.81 points across marks given to the 'Sentence structure and text organisation', 'Punctuation', and 'Composition and effect' domains for both Text 1 and Text 2 of the instrument (see Table 18 below). The writing assessment had a mean score of 17.93 points with an SD of 6.04 across the full sample (N=6,570). As can be observed in the density plot[8] in Figure 3, the shape of the distribution is bell-shaped and symmetrical and thus close to a normal distribution. The table of frequency distributions for the scale can be found in Appendix B. An analysis of the internal consistency of the 37 markers for Text 1 revealed an average Cronbach's Alpha of 0.83, with only one marker falling below 0.70 (0.69) points. For Text 2, the average Cronbach's Alpha was 0.83, with only two markers falling below 0.70 (0.63 and 0.67) points. Tables with the Cronbach's Alpha for each marker and the number of tests marked are included in Appendix E.

Table 18: Domain mark to total correlation for the writing assessment (TSWA)

Marking strand	Text 1	Text 2
Sentence structure and text organisation	0.83	0.83
Punctuation	0.79	0.80
Composition and effect	0.80	0.80

Figure 3: Density plot for the writing assessment (TSWA)



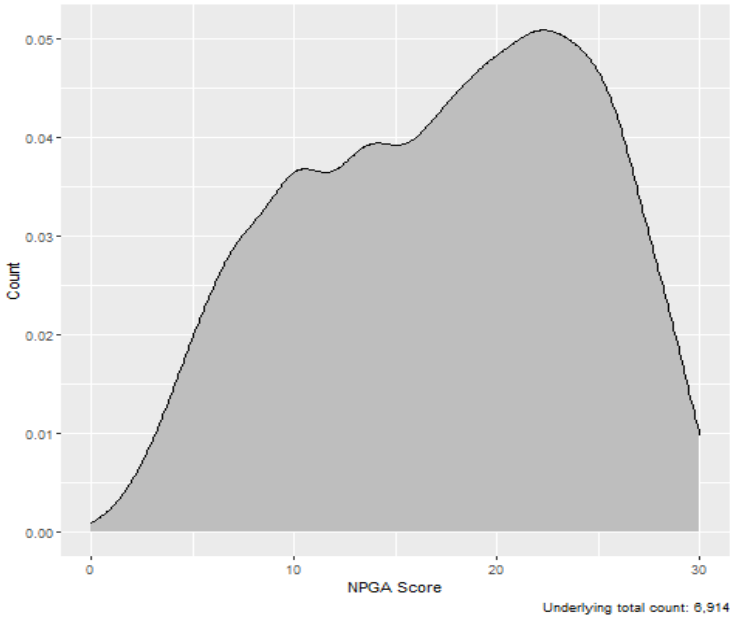
[8] A density plot shows a smoothed version of a frequency plot or histogram.

From the analyses of strand internal consistency, marker reliability and frequency distribution reveal that the functioning of the writing assessment (TSWA) was psychometrically adequate.

NPGA

The psychometric analysis of the 30-item grammar assessment (NPGA) indicated good internal consistency with a Cronbach's Alpha of 0.90. Moreover, the average item-total correlation was 0.45 points, and only one item fell below a 0.20 item-total correlation. The summated scale had a mean of 17.31 points with an SD 6.97 across the full sample (n=6,939). As can be observed in Figure 4, the shape of the distribution was slightly flattened, with a ceiling effect and slight left-side skew. As a rule of thumb, the psychometric literature considers that scales for which their lowest or highest scores account for 5% or more of respondents can be seen to present a considerable floor or ceiling effect, which can affect their measurement properties (Fisher Jr., 2007). Our scale's lowest point accounts for 0.06% and the highest score is just below 1% of the participants (0.73%). Consequently, we did not consider the ceiling effect to be large enough to be a significant threat to the measurement properties of the scale, or to warrant the need to carry out sensitivity checks. The table of frequency distributions for the scale can be found in Appendix B.

Figure 4: Density plot for the grammar assessment (NPGA)



From the analyses of item-internal consistency and frequency distribution, the functioning of the grammar assessment can be considered psychometrically adequate, with the caveat that there is a range restriction for those in the highest levels of the scale. Consequently, some respondents might have received higher scores if items with a higher level of difficulty were included in the assessment.

Correlation between prior attainment and the TSWA and NPGA

The correlations between the measures of Key Stage 2 attainment and the assessment instruments are presented in Table 19 below. Comparing the two Key Stage 2 attainment scores, the Key Stage 2 GPS score was more highly correlated with both the writing assessment score and grammar assessment score than the Key Stage 2 reading score. We therefore, included the Key Stage 2 GPS score in the models' covariates in order to control for prior attainment.

Table 19: Writing assessment and grammar assessment correlations with Key Stage 2 prior attainment measures

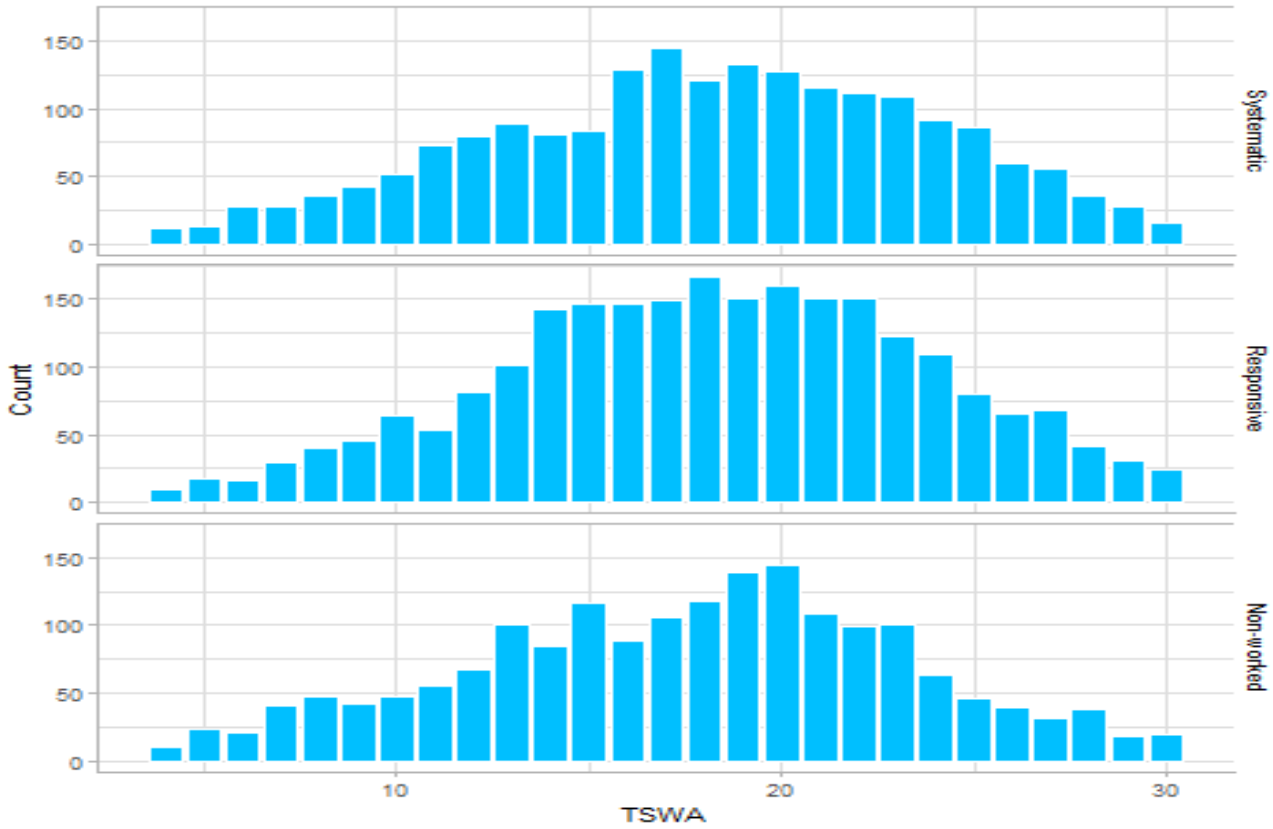
Key Stage 2 scores	Writing assessment			Grammar assessment		
	R	n	P-value	R	N	P-value
Key Stage 2 GPS score	0.56	6,297	< 0.001	0.79	6,642	< 0.001
Key Stage 2 reading score	0.50	6,292	< 0.001	0.67	6,637	< 0.001

Primary analysis

RQ1. What is the difference in writing composition of Year 7 pupils taught using the three different approaches, as measured by a bespoke TSWA?

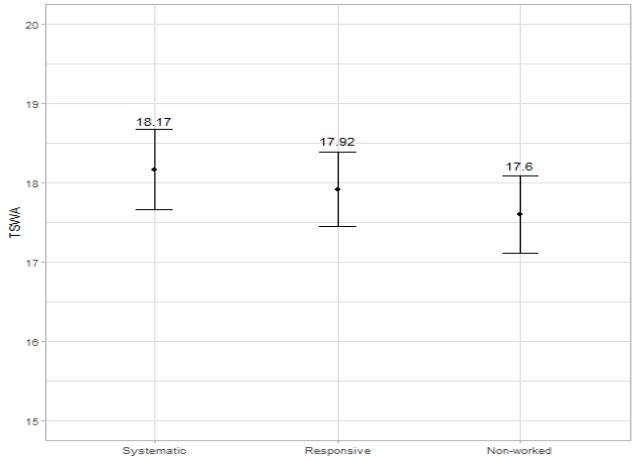
The primary analysis was undertaken on an ITT basis. The analysis (research question 1) investigated the impact of the different teaching approaches on writing composition measured using an end-of-trial TSWA. Figure 5 shows the frequency distributions of scores for each arm and Figure 6 shows the mean TSWA score and CIs for each of the different teaching approaches.

Figure 5: Frequency distributions for research question 1: Impact of the different teaching approaches choices on writing composition (TSWA score)



Note: Systematic worked 2,022, responsive worked 2,408, non-worked 1,867, total 6,297 counts (bars for TSWA values lower than 4 and higher than 30 points have been suppressed due to low counts).

Figure 6: Adjusted means and CIs for research question 1: Impact of the different teaching approaches choices on writing composition (TSWA)



A likelihood ratio test compared a null model to an alternative model testing the hypothesis that there is no difference in the mean scores between the three groups and through the difference in –2log-likelihood between a nested multi-level model (two levels: pupil; and teacher-class unit) with and without the two group dummies via a chi-squared test. To control for prior attainment, each pupils' Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test resulted in a chi-square of 3.79 with 2 degrees of freedom (p=0.150) and an ICC of 0.12. This indicates that the alternative model did not provide evidence of a significantly better fit to the data compared to the null model.

Table 20 presents unadjusted and adjusted[9] means and CIs for the primary outcome. The adjusted means show that the systematic worked arm teaching approach resulted in the highest mean scores on the writing assessment (18.17; 95% CI: 17.66, 18.67), followed by the responsive worked arm (17.92; 95% CI: 17.44, 18.41) and non-worked arm (17.60; 95% CI: 17.11, 18.08). Notwithstanding these mean differences, as noted above, the likelihood ratio test did not provide statistical evidence that the teaching approaches influenced pupils' writing assessment scores (see Table 20).

Although in the project study plan (Smith *et al.,* 2024), we said we would only run post-hoc tests where the likelihood ratio test provided evidence of a better fit, after consultation with the EEF we agreed to include this for the primary research question. Consequently, Table 21 below shows post-hoc pairwise differences between the means for each teaching approach and the equivalent in terms of effect size.

Table 20: Unadjusted and adjusted means and CIs for research question 1

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
TSWA (unadjusted)	2,098 (855)	17.86 (17.05, 18.66)	2,513 (740)	17.78 (17.00, 18.55)	1,934 (761)	17.18 (16.40, 17.95)
TSWA (adjusted)	2,022 (931)	18.17 (17.66, 18.67)	2,408 (845)	17.92 (17.44, 18.41)	1,867 (828)	17.60 (17.11, 18.08)

Table 21: Post-hoc pairwise writing assessment score differences for all pupils (research question 1)

Arm	Difference	Hedge's g	Lower Bound	Upper Bound	P-value
Non-worked – Systematic worked	-0.57	-0.07	-0.15	0.01	0.150 (p-value for the likelihood ratio test)*
Non-worked – Responsive worked	-0.33	-0.04	-0.12	0.04	
Systematic worked – Responsive worked	0.24	0.03	-0.05	0.11	

**Overall p-value for the likelihood ratio test (rather than the pairwise comparisons) shown as likelihood ratio test was not significant.*

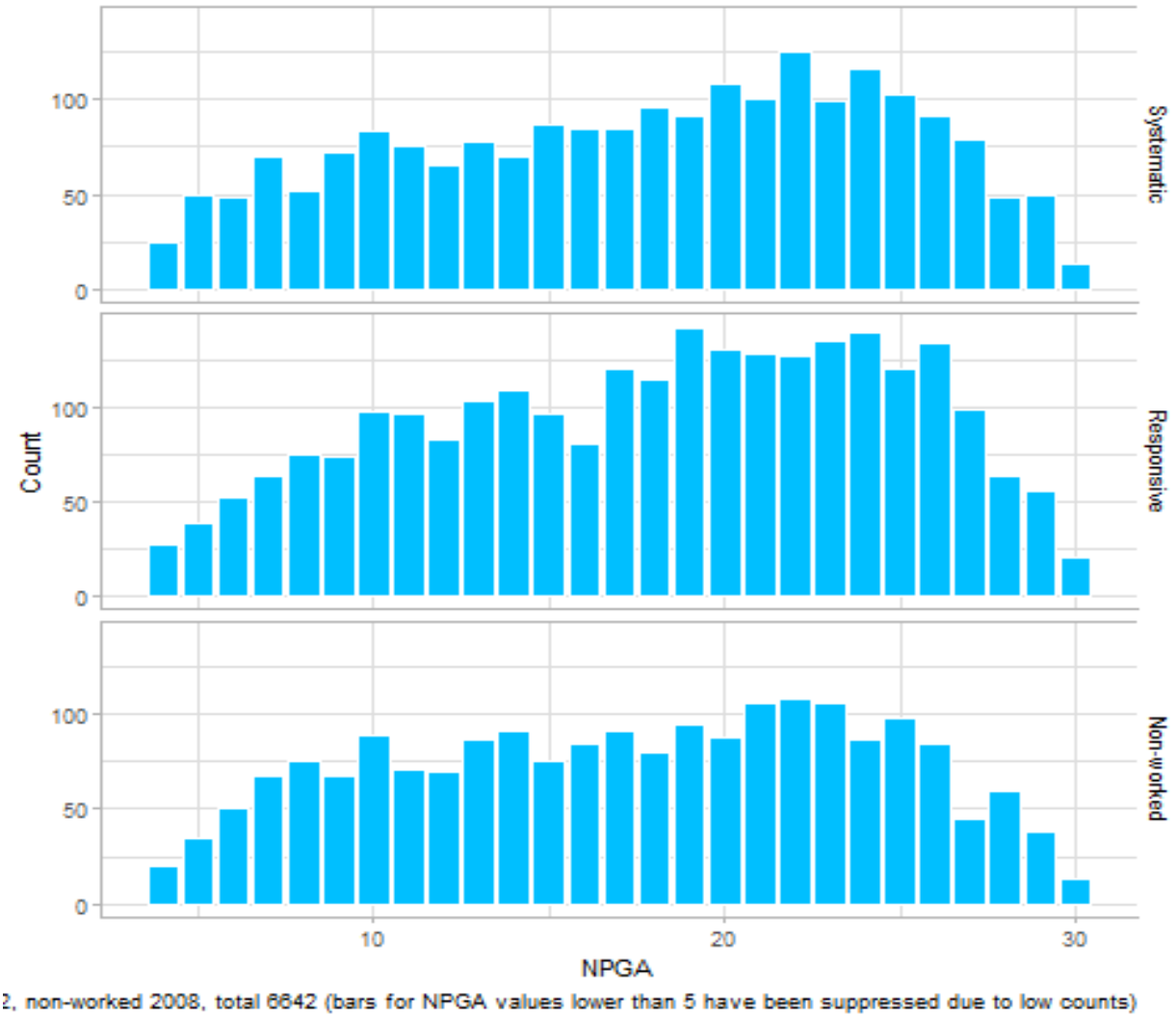
Secondary analysis

RQ2. What is the difference in knowledge of grammatical constructs of Year 7 pupils taught using the three different approaches, as measured by an end-of-block NPGA?

The secondary analysis (research question 2) used a different assessment to understand the impact of the teaching approaches on English grammar attainment, as measured by the NPGA administered at the end of the first block. Figure 7 shows the frequency distributions of scores for each arm and Figure 8 shows the NPGA mean score and CIs for each of the different teaching approaches.

[9]Unadjusted means are calculated from the raw data whereas adjusted means are taken from the model output and calculated for a pupil with average covariate values at an average school. Unadjusted means may be more prone to unseen biases in the school sample analysed, so for preference, adjusted means should be interpreted.

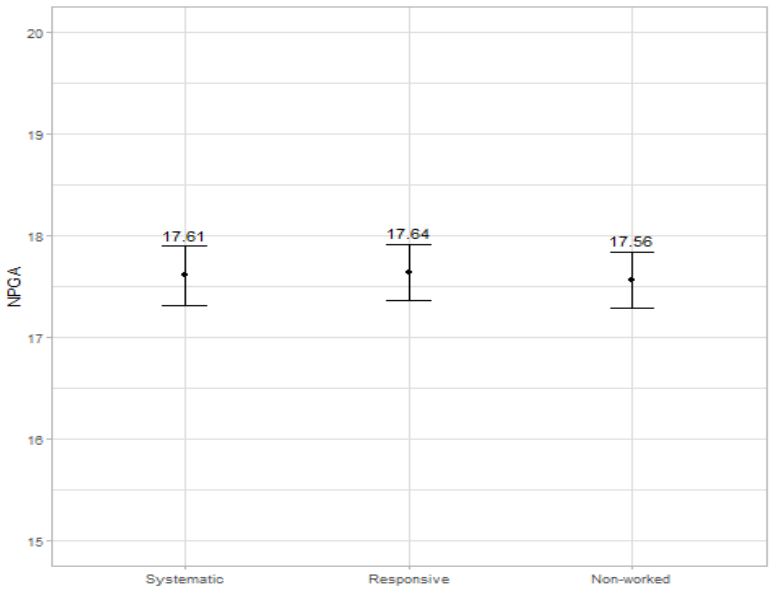
Figure 7: Frequency distributions for research question 2: Impact of the different teaching approaches choices on grammar attainment (NPGA score)



2, non-worked 2008, total 6642 (bars for NPGA values lower than 5 have been suppressed due to low counts)

Note: Systematic worked 2,092, responsive worked 2,542, non-worked 2,008, total 6,642 counts (bars for NPGA values lower than 4 points have been suppressed due to low counts).

Figure 8: Adjusted means and CIs for research question 2: Impact of the different teaching approaches choices on grammar attainment (NPGA)



A likelihood ratio test compared a null model to an alternative model testing the hypothesis that there is no difference in the mean scores between the three groups and through the difference in –2log-likelihood between a nested multi-level model

(two levels: pupil; and teacher-class unit) with and without the two group dummies via a chi-squared test. To control for prior attainment, each pupils' Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null model to the alternative model resulted in a likelihood ratio chi-square of 0.18 with 2 degrees of freedom (p=0.912, n=6,642) and an ICC of 0.04. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model.

Table 22 presents unadjusted and adjusted means and CIs for the secondary outcome. The adjusted means show that the responsive worked approach resulted in the highest mean scores (17.64; 95% CI: 17.37, 17.92), followed by the systematic worked approach (17.61; 95% CI: 17.31, 17.90) and non-worked approaches (17.56; 95% CI: 17.28, 17.84). Notwithstanding these mean differences, as noted above, the likelihood ratio test did not provide statistical evidence that the teaching approaches influenced pupils' writing assessment scores. As specified in the project study plan (Smith *et al.,* 2024), as the null hypothesis could not be rejected, no post-hoc tests were run to determine pairwise differences between the teaching approaches means nor their effect sizes (research question 2a is reported in the 'Subgroup analysis' section below).

Table 22: Unadjusted and adjusted means and CIs for research question 2

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
NPGA (unadjusted)	2,178 (775)	17.09 (16.07, 18.12)	2,656 (597)	17.38 (16.41, 18.35)	2,080 (615)	16.52 (15.56, 17.49)
NPGA (adjusted)	2,092 (861)	17.61 (17.31, 17.90)	2,542 (711)	17.64 (17.37, 17.92)	2,008 (687)	17.56 (17.28, 17.84)

RQ3. How do the differences in Year 7 pupils' writing composition vary by prior attainment (when measured by a bespoke TSWA)?

Research question 3 investigated whether the prior attainment level of the pupils interacted with the randomly allocated teaching approaches in regards to the writing assessment outcome. To control for prior attainment, each pupils' Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model, which included the interaction between prior attainment and the teaching approach dummies, resulted in a likelihood ratio chi-square of 2.98 with 2 degrees of freedom (p=0.226, n=6,297) and an ICC of 0.12. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there was no evidence that prior attainment had an effect on how the different teaching approaches impacted the pupils' writing assessment scores.

RQ4. How do the differences in Year 7 pupils writing composition vary by the number of sessions taught (when measured by a bespoke TSWA)?

Research question 4 investigated whether the number of teaching sessions to which pupils were exposed interacted with the teaching approaches in regards to writing assessment scores. To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (which included the interaction between the number of sessions and the teaching approach dummies) resulted in a likelihood ratio chi-square of 2.79 with 2 degrees of freedom (p=0.247, n=3,773) and an ICC of 0.09. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that being exposed to different numbers of sessions has an effect on how the different teaching approaches impacted the pupils' writing assessment scores.

Subgroup analyses

RQ1a (Analysis using FSM subsample): How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils? (relative to RQ1. What is the difference in writing composition of Year 7 pupils taught using the three different approaches, as measured by a bespoke TSWA?)

Following the logic model, which identifies that socio-economic disadvantage (i.e. FSM-eligibility) is systematically associated with lower prior attainment, and that use of worked examples is hypothesised to more greatly impact pupils with lower prior attainment, we estimated the model testing differences across teaching approach using the writing assessment (TSWA) with FSM-eligible pupils only.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (including the teaching approach dummies) resulted in a likelihood ratio chi-square of 8.23 with 2 degrees of freedom (p=0.016) and an ICC of 0.09. This indicated that the alternative model provided evidence of a significantly better fit to the data compared to the null model and thus, there might be a possible significant difference between some of the teaching approaches.

Table 23 presents unadjusted and adjusted means and CIs for the primary outcome and Table 24 presents the mean differences between teaching approach and their effect sizes calculated from the model. The adjusted means and differences show that the systematic worked examples teaching approach resulted in the highest TSWA mean scores (16.89; 95% CI: 16.18, 17.59) followed by the responsive worked examples (16.23; 95% CI: 15.59, 16.86) and non-worked examples approaches (15.89; 95% CI: 15.25, 16.53). As can be seen in Table 24, having corrected for familywise error using the Tukey method (Keselman and Rogan, 1977), none of the post-hoc tests assessing pairwise differences were significant, and thus sampling error could not be ruled out as the source of the significant log-likelihood result and of the observed differences. Consequently, there is no evidence that the writing composition of FSM pupils differed after being taught with the different teaching approaches when measured with the writing assessment.

Table 23: Unadjusted and adjusted means and CIs for research question 1a

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
TSWA (unadjusted FSM-pupils only)	436 (220)	16.87 (15.92, 17.81)	666 (245)	16.44 (15.58, 17.30)	484 (253)	16.03 (15.16, 16.89)
TSWA (adjusted FSM-pupils only)	421 (235)	16.89 (16.18, 17.59)	646 (265)	16.23 (15.59, 16.86)	473 (264)	15.89 (15.25, 16.53)

Table 24: Post-hoc pairwise writing assessment score differences for FSM pupils (research question 1a)

Arm	Difference	Hedge's g	Lower bound	Upper bound	P-value
Non-worked – Systematic worked	-1.00	-0.13	-0.24	-0.02	0.067
Non-worked – Responsive worked	-0.34	-0.04	-0.15	0.07	0.71
Systematic worked – Responsive worked	0.66	0.09	-0.03	0.20	0.292

RQ1a. (Analysis using FSM subgroup interaction): How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils? (relative to RQ1. What is the difference in writing composition of Year 7 pupils taught using the three different approaches, as measured by a bespoke TSWA?)

Further investigation of whether there were differences for FSM pupils was conducted by looking at whether the FSM status of the pupil interacted with the teaching approaches, which had been randomised in regards to the writing assessment (TSWA) outcome.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (which included the interaction between FSM and the teaching approach dummies) resulted in a likelihood ratio chi-square of 2.68 with 2 degrees of freedom (p=0.262, n=6,293) and an ICC of 0.12. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there was no evidence that the degree of change in TSWA scores between non-FSM and FSM pupils was different across the three teaching approaches (i.e. there was no evidence that the approaches were working differentially for non-FSM and FSM pupils).

The interaction analysis tests the difference between FSM and non-FSM pupils independently of whether there are mean score differences across the three teaching approaches. Consequently, the partially significant result in research question 1a (analysis using FSM subsample) is not contradictory to this result, which uses a subgroup interaction term (and the whole sample rather than FSM subsample). Research question 1a (using the FSM subsample) tested whether the TSWA scores that FSM pupils obtained were different across the teaching approaches.

RQ2a. (Analysis using FSM subsample): How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils? (relative to RQ2. What is the difference in knowledge of grammatical constructs of Year 7 pupils taught using the three different approaches, as measured by an end-of-block NPGA?)

Following the logic model, which identifies that socio-economic disadvantage (i.e. FSM-eligibility) is systematically associated with lower prior attainment, and that use of worked examples is hypothesised to more greatly impact pupils with lower prior attainment, similar to research question 1a, we estimated the model for differences across teaching approach, this time using the NPGA, and including FSM-eligible pupils only.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. Table 25 presents unadjusted and adjusted means and CIs for the secondary outcome. The adjusted means show that the responsive worked examples resulted in the highest grammar assessment mean scores (15.41; 95% CI: 14.97, 15.85) followed by the systematic worked examples (15.31; 95% CI: 14.80, 15.83) and non-worked examples approaches (15.22; 95% CI: 14.77, 15.67).

The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model resulted in a likelihood ratio chi-square of 0.50 with 2 degrees of freedom (p=0.779, n=1,675) and an ICC of 0.02. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there was no evidence that the FSM-eligibility of pupils changed the grammar assessment means resulting from the different teaching approaches.

Table 25: Unadjusted and adjusted means and CIs for research question 2a

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
NPGA (unadjusted FSM-pupils only)	460 (196)	15.25 (14.17, 16.34)	721 (190)	15.64 (14.68, 16.61)	556 (181)	15.05 (14.07, 16.03)
NPGA (adjusted FSM-pupils only)	437 (219)	15.31 (14.80, 15.83)	695 (216)	15.41 (14.97, 15.85)	543 (194)	15.22 (14.77, 15.67)

RQ2a. (Analysis using FSM subgroup interaction): How do any estimated differences vary between FSM-eligible and non-FSM-eligible pupils? (relative to RQ2. What is the difference in knowledge of grammatical constructs of Year 7 pupils taught using the three different approaches, as measured by an end-of-block NPGA?)

Similar to the approach to research question 1a, we carried out further investigation of whether there were differences for FSM pupils using an interaction model. This research question investigated whether the FSM status of the pupil interacted with the teaching approaches, which had been randomised as reflected by the NPGA outcome.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (which included the interaction between FSM and the teaching approach dummies) resulted in a likelihood ratio chi-square of 1.19 with 2 degrees of freedom (p=0.551, n=6,638) and an ICC of 0.04. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that the FSM status of pupils changes the NPGA scores obtained under the different teaching approaches.

RQ3a. (Analysis using EAL subsample): How do the differences in Year 7 EAL pupils writing composition vary by prior attainment when measured by a bespoke TSWA?

Research question 3a investigated whether the prior attainment of pupils changed the impact that the teaching approaches have on the scores obtained on the writing assessment when including only EAL pupils.

To control for prior attainment, each pupils' Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model, which included the interaction between prior attainment and the teaching approach dummies, resulted in a likelihood ratio chi-square of 4.45 with 2 degrees of freedom (p=0.108, n=1,121) and an ICC of 0.15. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that the prior attainment of EAL pupils affects the results obtained on the writing assessment when being exposed to the different teaching approaches.

RQ4a. (Analysis using FSM subsample): How do any estimated differences vary between FSM-eligible and non-FSM eligible pupils? (relative to RQ4. How do the differences in Year 7 pupils writing composition vary by the number of sessions taught, when measured by a bespoke TSWA?)

Research question 4a investigated whether the number of teaching sessions that the subgroup of FSM-eligible pupils were exposed to affected the writing assessment scores obtained under the different teaching approaches. To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The likelihood ratio test comparing the null nested

multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (which included the interaction between sessions and the teaching approach dummies) resulted in a likelihood ratio chi-square of 0.08 with 2 degrees of freedom (p=0.962, n=877) and an ICC of 0.08. This indicated that the alternative model provided no evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that the number of teaching sessions that FSM-eligible pupils were exposed to affected their scores on the writing assessment under the different teaching approaches.

Additional analyses and robustness checks

The writing assessment measures three aspects of writing: 'Sentence structure and text organisation'; 'Composition and effect'; and 'Punctuation'. The logic model hypothesises that the worked example approaches will support systematic improvement in pupils' **sophistication** of use of grammatical constructions, reflected in the 'Sentence structure and text organisation' score, and pupils' **control** of use of grammatical constructions. However, we do not expect the example approaches to have a differential impact on the 'Punctuation' score.

Therefore, two models were assessed in order to support an attribution to the treatment: i) omit the 'Punctuation' component of the TSWA (assess 'Sentence structure and text organisation' and 'Composition and effect' only); and ii) use 'Punctuation' only as the outcome. The rationale is that finding support for (i) and not (ii) further supports that the impact of the evaluation is related to the specific **intent** of the teaching approaches and not to a more general effect on writing **proficiency**. The results for these models are presented below.

RQ1. ('Sentence structure and text organisation' and 'Composition and effect'): What is the difference in writing composition of Year 7 pupils taught using the three different approaches, as measured by the 'Sentence structure and text organisation' and 'Composition and effect' strands of the TSWA?

This research question assessed whether the teaching approaches seemed to impact the score obtained on the 'Sentence structure and text organisation' and 'Composition and effect' strands of the writing assessment (TSWA) in different ways.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. Table 26 presents unadjusted and adjusted means and CIs for the 'Sentence structure and text organisation' and 'Composition and effect' strands. The adjusted means show that the systematic worked examples teaching approach resulted in the highest TSWA mean subtask scores (13.80; 95% CI: 13.42, 14.18) followed by the responsive worked examples (13.65; 95% CI: 13.29, 14.02) and non-worked examples approaches (13.44; 95% CI: 13.08, 13.81).

The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model, including the teaching approach dummies, resulted in a likelihood ratio chi-square of 2.49 with 2 degrees of freedom (p=0.288; =6,180) and an ICC of 0.12. This indicates that the alternative model did not provide evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that the teaching approaches differed in regard to the scores pupils obtained on the 'Sentence structure and text organisation' and 'Composition and effect' strands of the writing assessment.

Table 26: Unadjusted and adjusted means and CIs for research question 1 ('Sentence structure and text organisation' and 'Composition and effect' strands)

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
TSWA unadjusted subscores (sum 'Sentence structure and text organisation' + 'Composition and effect')	2,065 (888)	13.59 (12.99, 14.20)	2,479 (774)	13.52 (12.94, 14.10)	1,877 (818)	13.11 (12.54, 13.69)
TSWA adjusted subscores (sum 'Sentence structure and text organisation' + 'Composition and effect')	1,993 (960)	13.80 (13.42, 14.18)	2,376 (877)	13.65 (13.29, 14.02)	1,811 (884)	13.44 (13.08, 13.81)

RQ1. ('Punctuation'): What is the difference in writing composition of Year 7 pupils taught using the three different approaches, as measured by the 'Punctuation' strand of the TSWA?

This research question assessed whether the teaching approaches seemed to impact the score obtained on the 'Punctuation' strand of the TSWA in different ways.

To control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. Table 27 presents unadjusted and adjusted means and CIs for the primary outcome, 'Punctuation' strand. The adjusted means and differences show that the systematic worked examples (4.49; 95% CI: 4.37, 4.61) and responsive worked examples (4.49; 95% CI: 4.37, 4.60) teaching approaches resulted in equal TSWA subtask mean scores followed by a lower subtask score for the non-worked examples approach (4.43; 95% CI: 4.31, 4.55).

The likelihood ratio test comparing the null nested multi-level model (two levels: pupil; and teacher-class unit) to the alternative model (including the teaching approach dummies) resulted in a likelihood ratio chi-square of 0.74 with 2 degrees of freedom (p=0.692, n=6,180) and an ICC of 0.09. This indicates that the alternative model did not provide evidence of a significantly better fit to the data compared to the null model. Consequently, there is no evidence that the teaching approaches differed in regard to the scores pupils obtained on the 'Punctuation' strand of the writing assessment.

Table 27: Unadjusted and adjusted means and CIs for research question 1 ('Punctuation' strand)

	Means					
	Systematic worked		Responsive worked		Non-worked	
Outcome	N (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)
TSWA unadjusted subscore 'Punctuation'	2,065 (888)	4.41 (4.21, 4.60)	2,479 (774)	4.45 (4.26, 4.63)	1,877 (818)	4.32 (4.13, 4.50)
TSWA Adjusted subscore 'Punctuation'	1,993 (960)	4.49 (4.37, 4.61)	2,376 (877)	4.49 (4.37, 4.60)	1,811 (884)	4.43 (4.31, 4.55)

Missing data analysis

Table 28 quantifies the extent of missingness in the variables used in the research question models. As can be seen in Table 28 below, there are large proportions of missing data for the writing assessment (26%), grammar assessment (22%), and number of taught sessions (49%).

Table 28: Missing data for variables

Variable	Total n	Missing (n)	Missing (%)
Key Stage 2 GPS score	8,503	385	4%
Key Stage 2 reading score	8,494	394	4%
NPGA	6,914	1,974	22%
TSWA	6,545	2,343	26%
FSM	8,859	29	0.3%
Gender	8,859	29	0.3%
EAL	8,824	64	0.7%
Number of sessions taught	4,550	4,338	49%

Underlying counts: 8,888.

As indicated in the 'Methods' section above, we ran a two-level (pupil and teacher-class unit) multi-level logistic model in order to investigate the patterns of missing data for the writing assessment. Writing assessment missingness was a recode of writing assessment score as a dichotomous variable, with 1 indicating missing values and 0 indicating non-missing values. This included a fixed effect covariate for school, reflecting the stratified randomisation. Results for this regression (non-worked examples as the reference group) are presented in Table 29 below. Key Stage 2 attainment (Key Stage 2 GPS score), FSM, and EAL are predictive of writing assessment missingness. Given these statistically significant associations, we assumed that outcome data were MAR and included a model that examined research question 1 with the missing data imputed. This is described in the 'Methods' section above, and the results of this multiple imputed analysis of the primary outcome are included below.

Table 29: Regression for TSWA (writing assessment) missingness

TSWA	Estimate	Standard error	P-value
(Intercept)	0.89	0.28	0.001
Key Stage 2 GPS score	0.02	0.01	0.000
Key Stage 2 reading score	0.00	0.01	0.433
EAL	0.31	0.11	0.005
FSM	-0.54	0.08	0.000
Gender	-0.07	0.07	0.379
Systematic worked	-0.23	0.34	0.487
Non-worked	-0.19	0.34	0.571

Underlying counts: 8,888.

Following the project study plan, a similar analysis was carried out for missingness in Key Stage 2 prior attainment variables. Table 30 presents the multi-level regression (non-worked examples as the reference group) estimates for Key Stage 2 GPS missingness as a recode with 1 indicating missing and 0 indicating non-missing, including a fixed effect covariate for school, reflecting the stratified randomisation. Results indicate that only Key Stage 2 reading was a significant predictor of Key Stage 2 GPS missingness. Consequently, having higher Key Stage 2 reading scores was associated with having a missing value on Key Stage 2 GPS (missingness on prior attainment variables was due to the inability to match some pupils to the NPD dataset). Nevertheless, there was no evidence that missingness on Key Stage 2 GPS was associated with allocation to any of the teaching approaches.

Table 30: Regression for Key Stage 2 GPS missingness

Key Stage 2 GPS score	Estimate	Standard error	P-value
(Intercept)	4.84	1.47	0.001
TSWA	-0.08	0.09	0.387
Key Stage 2 reading score	0.10	0.05	0.029
EAL[a]	26.88	573931.28	1.000
FSM	-0.16	0.89	0.862
Gender	0.84	0.90	0.351
Systematic worked	0.13	0.92	0.885
Non-worked examples	0.73	1.16	0.531

Underlying counts: 8,888.

[a]The standard error for the EAL coefficient estimate was very high and should be disregarded. Not including EAL in the regression gave comparable results for the other coefficients.

Table 31 presents the multi-level logistic regression estimates for Key Stage 2 reading missingness as a recode with 1 indicating missing and 0 indicating non-missing, including a fixed effect covariate for school, reflecting the stratified randomisation. Results indicate that writing assessment score, Key Stage 2 GPS score, EAL, FSM, and gender were significant predictors of Key Stage 2 reading missingness. Specifically, having lower writing assessment scores, higher Key Stage 2 GPS scores, having EAL status, not having FSM status, and being female seemed to be associated with having a missing value on Key Stage 2 reading. There was no evidence that missingness on Key Stage 2 GPS score was associated with having been assigned to any of the teaching approaches. Given that missingness in the Key Stage 2 prior attainment variables occurred before randomisation, it is not to be considered a source of bias.

Table 31: Regression for Key Stage 2 reading missingness

Key Stage 2 reading score	Estimate	Standard error	P-value
(Intercept)	7.94	0.01	0.000
TSWA	-0.16	0.01	0.000
Key Stage 2 GPS score	0.32	0.01	0.000
EAL	-0.70	0.01	0.000
FSM	0.16	0.01	0.000
Gender	1.78	0.01	0.000
Systematic worked	-1.52	1.16	0.191
Non-worked	0.01	2.00	0.994

Underlying counts: 8,888.

Research question 1. Analysis with multiple imputed data

In order to assess the sensitivity of the findings to attrition, given the results presented in research question 1 above, we imputed missing values for the TSWA outcome using the multi-level predictive mean matching procedure, resulting in five plausible datasets (8,888 observations). As before, in order to control for prior attainment, each pupil's Key Stage 2 baseline measure was included in the likelihood ratio test as a covariate, alongside a fixed effect covariate for school, reflecting the stratified randomisation. The pooled results for the nested multi-level model (two levels: pupil; and teacher-class unit) using imputed data did not change the results obtained with missing data.

The likelihood ratio test comparing the null model to the alternative model (including the teaching approach dummies) resulted in a pooled F-test of 0.84 points (p=0.434). This indicates that the alternative model did not provide evidence of a significantly better fit to the data when compared to the null model when using multiple imputed data. Consequently, there is no evidence from this sensitivity analysis that the different teaching approaches had an impact on the pupils' writing assessment scores.

This sensitivity check provides important evidence about the robustness of the primary findings, in the context of high attrition (29%) after randomisation, and differential attrition across the three teaching approaches (26% for responsive worked arm, compared with 31% each for systematic worked and non-worked arms). Our missing data analysis established that pupils for whom no primary outcome (TSWA) data was provided tended to have lower previous attainment scores than those for whom data was provided. The imputed data analysis described above used these previous attainment scores, together with the other covariates included, to predict the missing TSWA scores and re-evaluate research question 1 in light of a complete dataset. Since the results of this analysis did not differ from those obtained with the complete case analysis, we do not think that the high attrition, or differential attrition across approaches, pose a threat to the results obtained through the complete case analysis.

Implementation and process evaluation results

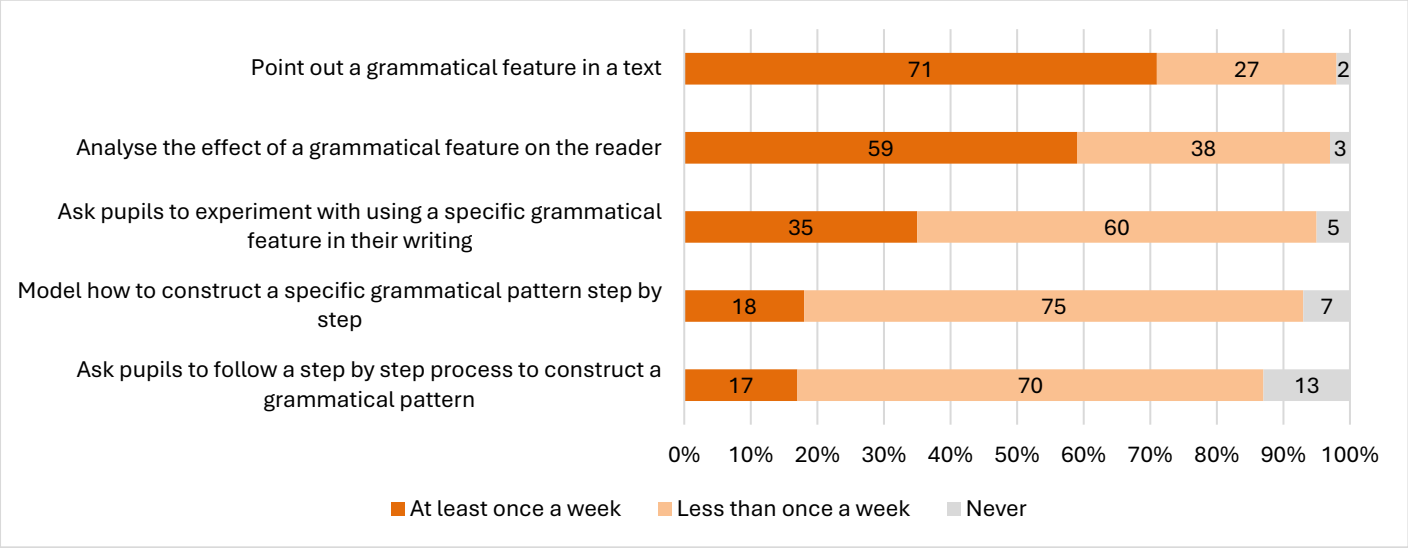
Context

Baseline survey data indicate that teacher participants had high levels of experience and specialism in English teaching. Almost all teachers primarily taught English during the 2023/2024 academic year (97%) and had completed teacher training in secondary English (91%). Most participants were experienced teachers: the median teaching experience was ten years, and only 10% of teachers were in their first two years of teaching. Most trial participants were classroom teachers (69%), with most of the remaining participants being head of English (13%) or deputy/assistant head of English (9%). The most common undergraduate degrees reported by teachers were English literature (45%) or a joint English language and literature degree (18%). This reflects other research indicating that English teachers tend to have a greater specialism in literature than language (Blake and Shortis, 2010).

Usual practice

The baseline survey asked teachers about their Year 7 grammar teaching to understand how the three grammar approaches compared with teachers' usual practice. More than half (61%) of teachers said they usually integrated grammar within other teaching content, in line with the responsive worked approach. Almost a third (29%) usually taught grammar as a separate activity (e.g. starter task), in line with the systematic worked and non-worked approaches. Very few teachers typically taught grammar in separate lessons (7%) or did not explicitly teach grammar (4%). Over half of teachers reported that they would identify grammar patterns in text and analyse the effect of such patterns on the reader 'at least once a week'. Less than half of teachers reported using elements of 'worked example' approaches to grammar 'at least once a week', by asking pupils to use a specific grammatical feature in writing (35%), modelling a step-by-step construction of a specific grammatical pattern (18%), or asking pupils to follow a step-by-step process to construct a grammatical pattern (17%). Figure 9 shows the self-reported frequency of these practices in teachers' usual Year 7 grammar teaching.

Figure 9: Frequency of teachers' use of specific strategies in grammar teaching



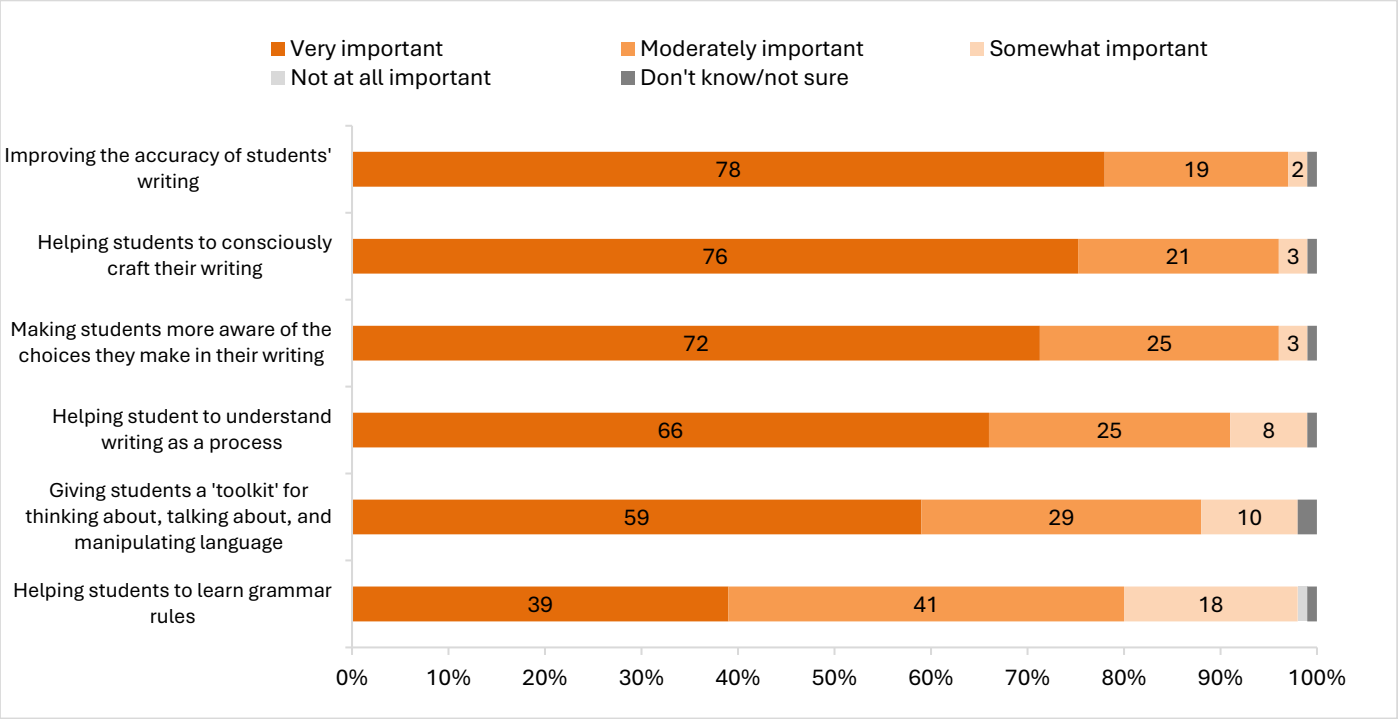
Data from baseline survey. On average, how often do you do the following in your lessons for a Year 7 English class? (N=202).

When teaching, teachers most commonly used examples from authentic texts[10] (90%), examples that they had written themselves (90%), or examples in resources from their colleagues/department (88%). It was less common for teachers to use examples from external curriculum resources (47%), from pupils in previous classes, (29%) or from GenAI (Generative Artificial Intelligence) (8%).

[10] 'Authentic texts' are texts, which are written for a general audience, rather than constructed specifically for teaching purposes, including novels, short stories, plays, and speeches.

Teachers were also asked about the importance of different uses of grammar in their Year 7 teaching (Figure 10). All the listed uses were considered 'very important' or 'moderately important' by most teachers (≥80%). The uses with the highest importance were improving the accuracy of writing (78% 'very important'), helping pupils to consciously craft their writing (76%), and making pupils more aware of the choices they make in their writing (72%). These matched the Key Stage 3 curriculum focus, emphasising pupils' use of grammar within writing, and awareness of writing choices, over the grammar rules emphasised in the Key Stage 2 curriculum.

Figure 10: Teachers' views on the importance of different aims for grammar teaching



Data from the baseline survey. In your Year 7 teaching, how important are the following uses of grammar? (N=202). Due to rounding, not all percentages sum to 100%.

Commenting on their usual approach to teaching creative and persuasive writing prior to the trial, teachers commented that they would typically use a guided writing approach, providing pupils with a writing frame or specific criteria for sentence and punctuation types, which they should include in their text. One example is provided below:

> *When we teach creative writing, we will say, for example, your first sentence has to have three words, your second sentence has to start an -ing word, your third sentence needs a simile or a metaphor.* (Head of English, Systematic worked approach)

Comparing trial choices with teachers' usual practice

Teachers reported using model texts to demonstrate good writing in their usual practice. Teachers suggested that the grammar approaches supplied in the trial differed from usual practice by emphasising a specific grammar pattern, rather than an overall model response, but were similar in focusing on consciously crafting language and the effect of these choices on the reader.

Through the endpoint survey, just over half of teachers (53%) agreed that their allocated teaching approach was like their usual teaching (responsive worked = 64%; systematic worked = 57%; non-worked = 33%). Aligned with these findings, teachers allocated to the non-worked approach were most likely to agree that this approach was new for pupils (60%). Around half of teachers allocated to the systematic worked approach reported this approach was new to their pupils (49%), compared to around a third of teachers allocated to the responsive worked approach (36%).

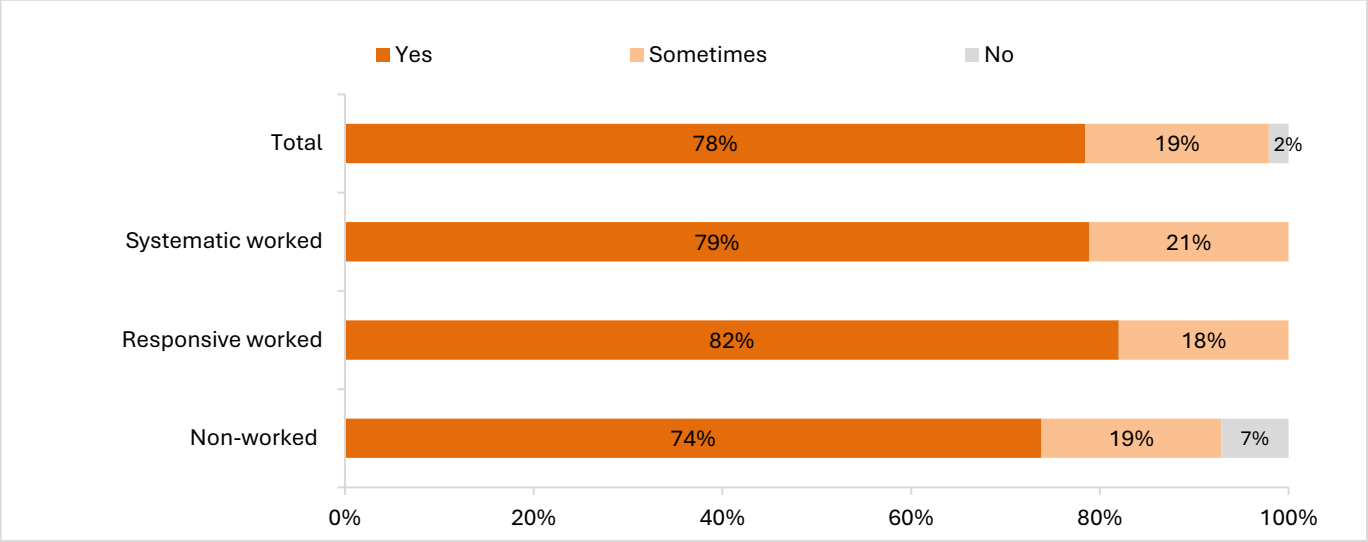
Focus groups with teachers in case study schools provided further information on how the grammar approaches differed from their usual approach to Year 7 grammar teaching. Before the trial, only one case study school reported following a

teaching scheme, which aimed to develop grammar. Teachers in the other case study schools felt the grammar approaches provided through the trial provided an explicit and structured approach to teaching grammar. This was a difference in practice compared to their usual practice of responding primarily to errors in pupils' grammar understanding, identified through pupil writing. In contrast, the grammar examples introduced a range of grammar patterns to expand pupils' repertoire of writing choices.

Adherence

In the endpoint survey, teachers were asked whether they had followed their allocated approach in their teaching sessions (see Figure 11). These self-reports indicate that most teachers felt they were using their allocated approach: 78% of teachers said that they had followed their allocated approach in their teaching sessions; about a fifth (19%) said that they had followed their approach 'sometimes'; while only 2% of teachers said that they had not followed their approach. Adherence was similar for each of the teaching approaches.

Figure 11: Proportion of teachers who reported adherence to their allocated teaching approach



Data from the endpoint survey (N=144). Research question 1. Your allocated teaching approach was X. Overall, were you able to follow this approach in the grammar sessions you taught?

The teachers who indicated that they had sometimes, or had not, followed their allocated approach (n=31) were asked to indicate the reasons from a closed list. The most common reasons were because their approach did not fit with their teaching scheme (n=17), took too long (n=17), or was not accessible for their class (n=13).

Teachers in case study schools reported that they had adhered to the grammar approach they were allocated to and were conscious not to discuss their approach with colleagues so as to not 'contaminate' their implementation.

Responsiveness

Teacher Responsiveness

The endpoint survey asked teachers about their views and experiences of their allocated approach. Responses show that overall, about two-thirds of teachers responded positively to their allocated approach, with some variance across items. In general, the two worked example approaches (systematic worked and responsive worked) had a higher proportion of positive responses than the non-worked approach.

Around two-thirds of teachers (65%) agreed or strongly agreed that their allocated teaching approach was easy to use. A smaller proportion (55%) agreed/strongly agreed that they had enjoyed using their allocated teaching approach (however, note that 26% neither agreed nor disagreed). Teachers allocated to the worked approaches were slightly more likely to agree that they had enjoyed using their approach (60% of systematic worked teachers, 55% of responsive worked teachers) than the non-worked approach (48%).

These findings are echoed by teachers in case study schools. Teachers allocated to the systematic worked approach reported that they liked this approach for the structure and routine it provided to lessons. They also liked the structure of the sessions, in which pupils identified the grammar pattern and explored the effect of this on the reader. Teachers felt the conversations in which they discussed how the author had used the pattern was particularly valuable for supporting pupils to replicate the pattern in their own writing and achieve the effect they wanted their own text to have on a reader. Teachers allocated to the responsive worked approach reported that they found this approach somewhat disjointed from the rest of the lesson at the start of the trial. However, as they had become more familiar with the approach, the grammar sessions had become better woven into lessons and teachers gained confidence in applying the session to an element of the main lesson. For these reasons, there was the perception that the responsive worked approach would be the most efficacious because the grammar session was built into the lesson and gave pupils the opportunity to connect grammar with their wider writing, rather than it being an isolated, decontextualised task. In comparison, teachers allocated to the non-worked approach felt that sessions were disjointed and unrelated to the main lesson content. Across the case study schools, there was an uncertainty of the effectiveness of the non-worked approach as it lacked the key step in learning of practicing and applying what has been learned. Across case study schools more widely, teachers tended to report however that they had enjoyed delivering the grammar sessions. They also enjoyed being part of a Teacher Choices trial for the perceived benefits brought about for pupils and the potential application of one or a combination of the approaches to departmental schemes of learning (discussed further in the 'Perceived outcomes' section). Some case study teachers whose specialism was English language or linguistics reported high enjoyment, attributing this to more confidence in understanding and analysis grammar patterns, compared to English literature specialists, who lacked confidence in their own grammatical knowledge at the start of the trial.

Teachers were asked in the survey to indicate the extent to which they agreed that the focus of the grammar tasks aligned with their Year 7 curriculum/topics. Overall, a slightly larger proportion agreed that the focus on noun phrases in narrative fiction aligned with their curriculum/topics (65%; responsive worked 72%; systematic worked 63%; non-worked 59%) than the focus on clauses/sentences in persuasive writing (61%; responsive worked 64%; systematic worked 67%; non-worked 53%).

The findings from interviews with teachers in case study schools provide further insights into the extent to which the grammar tasks aligned with their curriculum/topic timeline. The first part of the trial (Summer Term 1) aligned well for schools that were covering creative writing, while it was a mismatch for other schools that had covered creative writing in the Spring Term. Teachers found the persuasive clauses/sentences (Summer Term 2) more difficult to align, as all schools were teaching Shakespeare. Teachers reported that these sessions often felt disjointed from the remainder of the lesson and 'shoehorned in'. This was particularly challenging for teachers in the responsive worked approach, who were expected to weave the grammar session into their lesson content.

Focus groups with teachers in case study schools offered additional insight about how teachers felt about the grammar patterns. Across all approaches, the main challenge teachers faced in delivering the grammar sessions was to not name the patterns. While the guidance asked teachers not to 'over dwell' on grammatical terminology,[11] and some teachers considered the formal pattern names would be inaccessible to their pupils, many teachers felt the absence of a label restricted pupils' retrieval of the patterns and future application.

> *I think that's a problem…they've got a temptation of feature spotting, giving it a name. I feel like teachers do as well, like I keep wanting to give it a name, like say 'can you use the one [pattern]' and give it a specific name so that they can pick it up easily…it's a way of categorising it for them, which I think they do need.* (Teacher, Responsive worked approach)

Linked with this, teachers found it difficult to 'hold back' and not prompt or recap elements of grammar, which pupils should have prior knowledge of, such as nouns or adjectives, which they felt would then help them in identifying the pattern. The

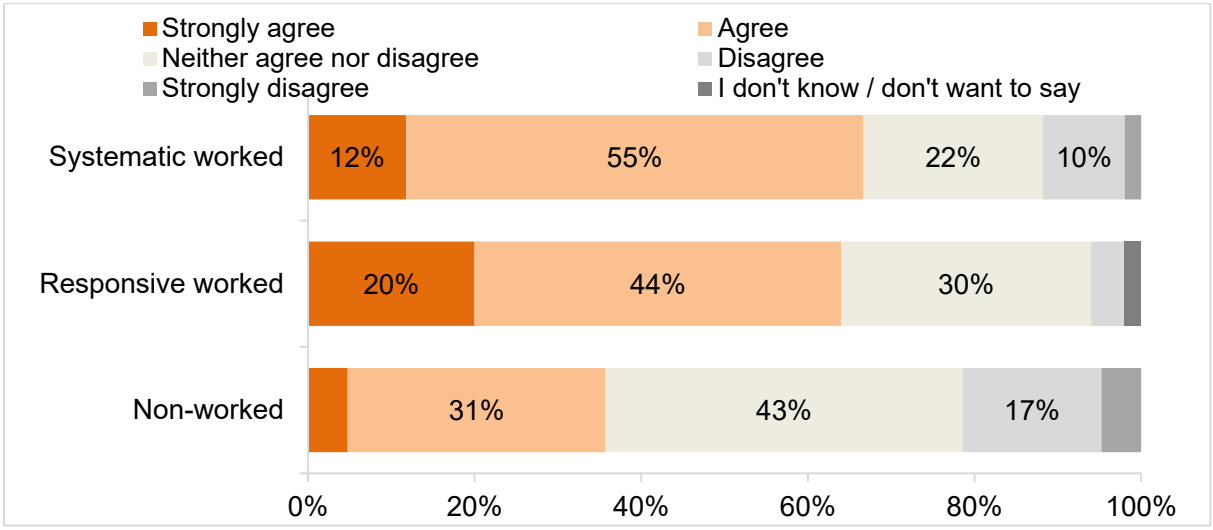
[11] The guidance stated: 'Don't…over dwell on the grammatical terminology or abstract understanding of the structure'.

impact of pupils' prior grammatical knowledge is covered within the 'Responsiveness' section. As discussed above, case study teachers valued the grammar sessions but often struggled to deliver these within the suggested 15 minutes.

Teacher perceptions of pupil responsiveness

A small majority of teachers (57%) perceived that, compared with their usual grammar teaching, their pupils had engaged well with the allocated teaching approach (however, note that 31% neither agreed nor disagreed). Responses were more positive from teachers in the worked example approaches compared with the non-worked approach. See Figure 12 below for details.

Figure 12: Teacher perceptions of pupil engagement



Data from endpoint survey (N=141). Compared with other grammar teaching, my pupils have engaged well with the allocated teaching approach. Percentages may not add to 100% due to rounding.

Focus group teachers shared their views on how pupils had engaged with the grammar approaches. Teachers reported that pupils had engaged well with the grammar sessions, for example, they enjoyed the opportunity to write creatively. Aligned with findings related to teacher responsiveness, pupils were also perceived to respond well to the routine that the grammar sessions created. Teachers suggested the similarity to Year 6 literacy content created a bridge between primary and secondary school, and pupils' prior familiarity with the terms and techniques presented gave pupils confidence to engage with the examples. For this reason, teachers also reported that pupils who had a positive previous experience with the Year 6 national curriculum assessments (i.e. mainly higher attaining pupils) had enjoyed the trial grammar assessment, given the similarity to their Year 6 assessments.

However, there was a consensus among case study teachers that higher attaining pupils had been able to access and therefore, engage with the grammar patterns better than lower attaining pupils, who did not have the level of grammatical knowledge required to confidently engage with grammar examples. Teachers reported that while pupils could identify language devices, such as similes and metaphors, or recognise the creative or persuasive nature of the text, they struggled to identify the grammar pattern, which was an unfamiliar task for them. Teachers reported that patterns, which were too difficult for pupils to identify and engage with had led to engagement and behavioural challenges. Teachers also reported that sometimes pupils became so focused on replicating the grammar pattern that their sentences stopped making sense and their usual creativity was lost. There were also cases where higher ability pupils were perceived to struggle with the sessions too. For example, with the complex narrative patterns teachers found that sometimes pupils struggled to synthesise what they had identified within the pattern and replicate this in their own writing.

> *When I was teaching them post-modifying with prepositional phrases, they get so obsessed with the prepositions that they lost track of the point of it. Sometimes, they are too complicated, there is a cognitive load issue. So, they will get the prepositions and make sure they have included those, but then they can't remember to do the other bit as well.* (Teacher, Responsive worked approach)

Although pupils had liked the chance to express themselves creatively using the descriptive noun phrase patterns, teachers perceived the persuasive patterns had been easier for pupils to grasp, not only because the patterns were less complex but also because most of the examples came from speeches by Greta Thunberg or Malala Yousafzai who pupils knew of (although as noted in the 'Choice enactment and fidelity' section, this context sometimes had to be provided by teachers), compared to the narrative phrases, which came from a wider range of largely unfamiliar books and authors.

Pupil perceptions of responsiveness

To further understand pupils' responses to the grammar examples, focus groups with pupils from the worked example approaches (six focus groups, four from systematic worked examples, and two from responsive worked examples, each including about five to six pupils) included an activity for them to share what they had liked or enjoyed about the grammar tasks (which were written onto stars) and what they had disliked or felt could be improved (which were written onto wishes). Pupils' comments fell into four main themes, discussed below.

Grammar patterns

Pupils allocated to the systematic worked approach fed back that they liked learning new grammar patterns to use in their writing. Linked with the outcomes that pupils recognised (reported in the 'Perceived outcomes' section), they liked that the inclusion of grammar patterns helped to improve their writing through making it more exciting and interesting. They described the patterns as 'fun to use' and liked the chance to be creative. Although some pupils had found the patterns 'easy' to understand, others liked the challenge they presented. Pupils would have liked to have been taught even more grammar patterns, which they could use in their writing.

Pupils in the responsive worked approach of the trial reported that they would have liked to be taught the name of each of the grammar patterns, so they had a label to attach to each structure.

Text examples

Pupils liked the amount and range of examples that they were presented with over the course of the trial. However, experiences and views around the number of examples presented per lesson differed. Some pupils reported that they had found it helpful to see several examples of each of the grammar patterns. This suggests that teachers had presented the authentic text model, as well as the 'further examples' from the example bank.

> *I think it's helpful because if you don't understand the first grammatical pattern, you can see the rest of the examples*. (Pupil, Responsive worked approach)

In comparison, other pupils felt it would have been helpful to see more than one example per pattern to increase familiarity, suggesting that some teachers did not draw upon the additional examples provided for each grammar pattern.

Pupils liked being introduced to the grammar patterns through authentic texts as these were 'real' examples. They would have liked examples from '*more relatable texts'*, such as class texts or their own reading.

> *I think it can be quite helpful [to see examples from authentic texts], because you can see the grammatical pattern actually being used from a really good author, how they would use it…and they've obviously got to get readers hooked into their story, and we can use that to get readers hooked into our stories*. (Pupil, Responsive worked approach)

Session focus

Pupils allocated to responsive worked and systematic worked approaches liked the focused steps of the grammar sessions, in which they were required to identify the pattern being shown, discuss the use and impact of this pattern within the example texts in pairs or as a class, then replicate it in their own piece of writing. For example, one pupil group appreciated that their teacher had included a picture (such as of a mythical creature, as observed during one lesson), which pupils were required to describe using the grammar pattern. Pupils had enjoyed the opportunity to share with the class what they had written and to hear other pupils' ideas and interpretations but would have liked more time for such sharing following independent writing.

Some pupils allocated to the systematic worked and responsive worked approaches fed back that, in addition to the class discussion about the grammar pattern, they would have also liked their teacher to show how to use the pattern in writing through constructing an example together as a class. This notes a difference in teacher practice, as working through an example was listed as an element of the session within the teacher guidance for these approaches.

Timing

Pupils allocated to the systematic worked approach had liked that the grammar tasks took place as a lesson starter. They said that completing the grammar tasks first helped them to prepare for the lesson and meant the pattern they had reviewed and practiced was fresh in their minds to use during any writing tasks in the main lesson. Pupils also reported that they had found the grammar tasks to be a helpful way of recapping literacy skills they had covered in primary school. This comment aligns with findings from the teacher focus groups, in which teachers reported that the focus on grammar acted as a bridge between primary and secondary English given its strong focus in the Key Stage 2 curriculum and in Year 6 SATS.

Pupils expressed a range of wishes in relation to the timing of the grammar tasks. Pupils across the systematic worked and responsive worked approaches of the trial would have appreciated more time within the sessions to practice including the grammar patterns within their own writing. Pupils in some groups would have also liked more time to talk about the examples, rather than having to write in silence, however, this wish reflects teacher practice rather than guidance on delivery of the sessions. Pupils had two suggestions for how the timing of the grammar sessions could be changed to incorporate these wishes. The first session was to have one lesson solely dedicated to grammar per week:

> *One lesson a week on them [would be better]. We still need them [grammar patterns], it's just if there are too many of them, it interrupts the learning of what we are actually focussed on.* (Pupil, Responsive worked approach)

The second suggestion was to have longer sessions focusing on each grammar pattern, with suggestions of 20 to 30 minutes as opposed to 15 minutes. They felt this would be helpful to allow additional time to understand and practice the grammar patterns. Pupils would have also liked additional time to revisit grammar patterns covered in previous lessons.

Choice enactment and fidelity

The endpoint survey asked teachers about their teaching of grammar examples to ascertain fidelity to their allocated teaching approach. To encourage an honest description of teachers' practice during the trial, the same statements were shown to all teachers, regardless of their allocated approach. Therefore, in some cases, indicating 'no sessions' represents fidelity to the intended approach, whereas in others, indicating 'all sessions' represents fidelity. Table 32 below displays the percentage of teachers who adhered to specific elements of their allocated approach as set out in the teacher guidance. Ticks indicate a feature included in the teacher guidance for that choice. Crosses indicate a feature, which was contraindicated in the teacher guidance for that choice. Each cell displays percentages of teachers who did, or did not, implement a feature in their grammar teaching sessions. Cells where less than 70% of teachers indicated fidelity are shaded.

Table 32: Teachers' fidelity to their allocated grammar approach

Grammar approach	Responsive worked		Systematic worked		Non-worked	
	All or most sessions	Half or fewer sessions	All or most sessions	Half or fewer sessions	All or most sessions	Half or fewer sessions
Taught grammar session as separate activity	✗ 76%	✓ 24%	✓ 82%	✗ 18%	✓ 88%	✗ 12%
Used grammar example from authentic texts	✓ 72%	✗ 28%	✓ 82%	✗ 18%	✓ 81%	✗ 19%

Provided multiple examples of grammar patterns	✓ 80%	✗ 20%	✓ 82%	✗ 18%	✓ 74%	✗ 26%
Talked about effect of grammar choices on the reader	✓ 82%	✗ 18%	✓ 84%	✗ 16%	✓ 83%	✗ 17%
Showed pupils step-by-step construction of grammar pattern	✓ 76%	✗ 24%	✓ 78%	✗ 22%	✗ 43%	✓ 57%
Asked pupils to write text including grammar pattern	✓ 88%	✗ 12%	✓ 84%	✗ 16%	✗ 41%	✓ 60%
Asked pupils to consider effect of their text on reader	✓ 64%	✗ 36%	✓ 77%	✗ 24%	✗ 60%	✓ 41%

Cells where less than 70% of teachers indicated fidelity are shaded.

Overall, Table 32 shows there were three key challenges for fidelity:

- It appears that the responsive worked approach was taught more similarly to the systematic worked approach than the guidance stipulated. Over three-quarters of teachers (76%) in the responsive worked group indicated they taught 'all or most' sessions as a separate activity in the lesson, however, the intention was that the responsive worked group would integrate the grammar sessions to the lesson at appropriate points.
- There was greater teaching of a step-by-step breakdown and asking pupil construction of text in the non-worked approach than the guidance intended. Focus group teachers commented that it was very unusual and difficult for them to share a model text and then hold back from encouraging pupils to write a similar text. As step-by-step breakdown and pupil construction are the key differences between the worked and non-worked example approaches, this means that the non-worked approach was probably implemented more similarly to the worked approaches than intended.
- Finally, while asking pupils to consider the effect of the text on the reader was meant to be a key feature of the responsive worked approach, less than two-thirds of teachers reported doing this in 'all or most' sessions (64%). Almost as many teachers in the non-worked example group reported doing this in 'all or most' sessions (60%), however, this was not a feature of this approach in the guidance.

Teachers in the systematic worked approach demonstrated good fidelity to their guidance with more than three-quarters of the teachers reporting that the specified features were used in 'all or most' lessons.

Observed teaching in case study schools showed higher fidelity to each approach, compared to the survey responses. For example, in observations of non-worked examples, no teachers showed a step-by-step construction or asked pupils to write text including a grammar pattern. In observations of responsive worked examples, the grammar pattern was integrated into the overall lesson content. As the case study schools did not aim to be representative, only provide a single snapshot of teaching, and the presence of an observer may have influenced teachers' behaviour, the survey findings are considered a more accurate reflection of achieved fidelity.

Variation and adaptation in the use of grammar examples

The endpoint survey asked teachers if they had made any adaptations to their allocated teaching approach, compared to the teacher guidance. Overall, the most common adaptation was to increase the session length beyond 15 minutes (38%: systematic worked 43%; responsive worked 38%; non-worked 31%). The survey findings showed that smaller proportions of teachers shortened the length of grammar lessons to less than ten minutes (26%: non-worked 31%; systematic worked 26%; responsive worked 22%). The findings from the teacher focus groups with teachers allocated to the systematic worked and responsive worked approaches add insight to these survey findings. They reported that the time taken for pupils to understand and process the grammar pattern being explored, then use it in their own writing and consider the effect on the reader, was a longer task than anticipated. As explained by one teacher:

> *Some of them would have been able to tell you the little bits of grammar independently but to the synthesise 'I have this knowledge and I'm going to turn it into that creative thing', that has been a difficult step I think for a lot of our kids, which has meant it has rarely been a 15-minute starter because of the processing that they've got to do, and the base knowledge that we would assume that they have that actually they haven't.* (Head of English, Systematic worked approach)

The different approaches had different guidance for the frequency of sessions. Teachers using the systematic worked and non-worked approaches were asked to use grammar examples in two lessons a week, for 15 minutes, to ensure regular spacing and repetition of the grammar pattern. Teachers allocated to the responsive worked approach were asked to use grammar examples 20 times during the trial period, whenever in the lesson sequence they felt appropriate, to allow a more naturalistic use of the examples. For this reason, only teachers allocated to systematic worked and non-worked approaches were asked about adaptations made to the timing of their delivery. The most common adaptation was to teach sessions with different spacing (i.e. not twice a week), which was more common within the systematic worked approach (43%) than the non-worked approach (24%).

Similarly, teachers using the systematic worked and non-worked approaches were asked to teach each grammar pattern twice. Fewer teachers indicated that they had spent fewer than two sessions on a grammar pattern, yet again this was more common in the systematic worked approach (28%) than non-worked approach (10%). Just 10% of teachers in both the systematic worked and non-worked approaches indicated that they had spent more than two sessions on a grammar pattern.

Around a fifth of all teachers (21%, n=30) recorded that they had not made any adaptations. Teachers allocated to the responsive worked approach were most likely to report this (n=19), compared to smaller numbers of teachers allocated to the non-worked (n=7) and systematic worked (n=4) approaches.

Teacher focus groups provided further details of how teachers had used their allocated approach. Most commonly, teachers reported that they had tailored their teaching to the attainment of their class and considered knowledge progression within and across sessions.

Within the non-worked approach, case study teachers adapted the complexity of vocabulary in the provided text examples, to provide additional support or challenge. Teachers using the worked approaches (systematic and responsive) reported that their higher attaining pupils had found it easier to grasp and replicate the grammar patterns in their own writing. Some provided additional scaffolding for lower attaining pupils, for example, using writing frames, which included sentence openers and the structure of the grammar pattern, so pupils used gap-fill tasks to create their own text. Several teachers of lower attainment classes reported that they had not used the most complex grammar patterns in the example bank, because the ceiling of these pupils' grammar skills was in identifying adjectives, nouns, and verbs and it would therefore, not have been beneficial to introduce more complex structures.

Teachers in one school (across all approaches) used a knowledge check of foundational grammar terms (e.g. nouns and adjectives) before presenting the example, to support pupils to engage with the more complex grammar patterns.

Teachers reported that when presenting authentic text examples, particularly those from Greta Thunberg and Malala Yousafzai, it was important to provide pupils with details around who the speakers were, and what they advocated for, so that pupils understood the context of the speeches. Teachers felt that this supported pupils to write their own examples, but also that it was inherently important for pupils to know about these speakers.

Another teacher in the systematic worked approach described incorporating retrieval practice to link across grammar patterns, namely, revisiting past grammar patterns for pupils to include when editing their work. They reported that they would make retrieval practice a key feature of teaching grammar patterns after the trial, systematically revisiting previous grammar patterns to support pupils in building on their learning.

More generally, several teachers had concerns over the extent to which they could make adaptations to the lessons. There was a feeling of needing to deliver sessions exactly as described in the optional materials, to ensure their practice mirrored

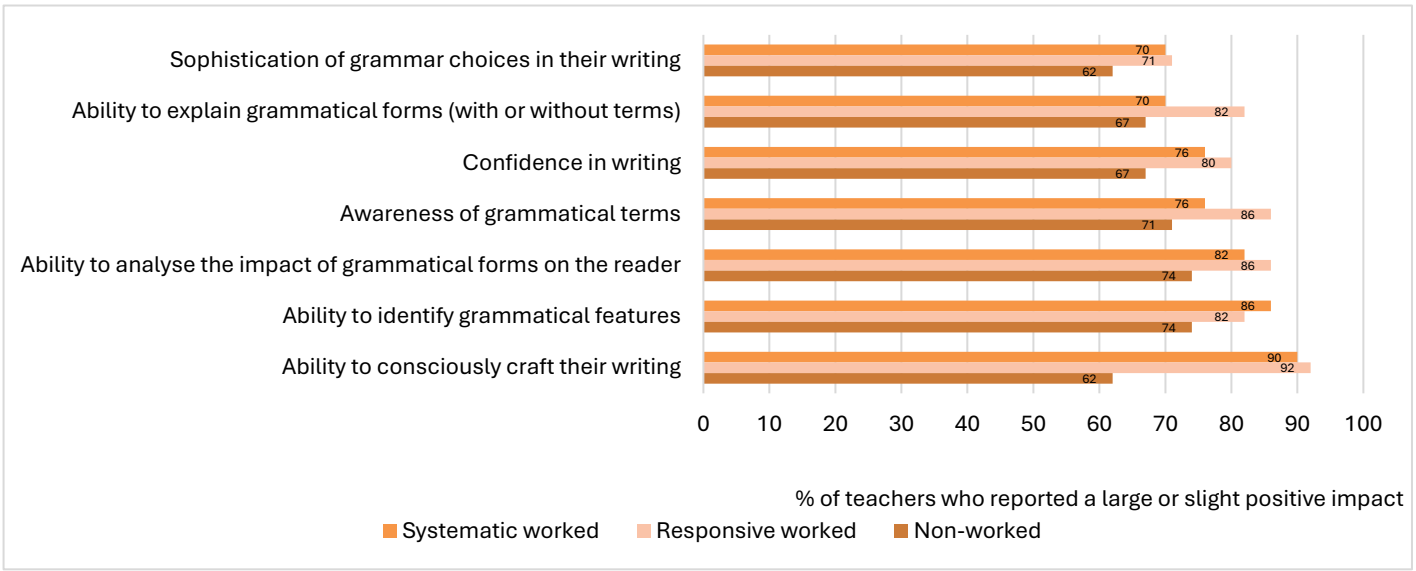
that of other schools, so fair comparisons could be made. These teachers wanted either more prescription to ensure their delivery was as expected, or guidance on how to make acceptable adaptations. For example, several teachers reported that they did not provide simplified language, different grammar patterns, or writing frames for their lower attaining groups, because they wanted to follow the trial guidance. This is an important consideration for the messaging across Teacher Choices trials, which expect teachers to use their allocated choice in a way which works for their class, compared with programmatic trials where fidelity is often more prescriptive.

Perceived outcomes

Perceived outcomes for pupils

Through the endpoint survey, teachers indicated the extent to which they perceived the grammar approaches to impact upon a range of pupils' skills (see Figure 13). Across all three approaches and across all seven items, over two-thirds of teachers perceived there to be a positive impact on pupils, although more teachers reported a 'slight' impact rather than a 'large' positive impact. Across all items, more teachers allocated to the worked approaches of the trial perceived their approach to have a positive impact on pupils (responsive worked 71–92%; systematic worked 70–90%; non-worked 62–74%). The worked example approaches were thought to have the greatest impact on pupils' ability to consciously craft their writing (responsive worked 92%, systematic worked 90%, compared with 62% for non-worked). The non-worked approach was thought to have the greatest impact on pupils' ability to analyse the impact of grammatical forms on the reader (74%) and identify grammatical features (74%), as might be expected from the session focus. However, even in these areas, the worked example approaches had higher perceptions of positive impact (responsive worked 86% and 82%, respectively; systematic worked 82% and 96%, respectively).

Figure 13: Teachers' views on pupil impacts



Data from the endpoint survey. What impact have you noticed on pupils? (Responsive worked N=49; systematic worked N=50; non-worked N=42). Due to rounding, not all percentages sum to 100%.

Teachers who participated in focus groups within case study schools shared the outcomes they had recognised for their pupils resulting from the grammar sessions. Broadly, teachers acknowledged the importance of explicitly teaching grammar to pupils for underpinning high-quality writing and analysis of texts.

Data from teacher focus groups and lesson observations suggest that the responsive worked and systematic worked approaches succeeded in achieving proximal use of the grammar patterns in pupils' own writing. When focused on writing a few sentences, which included the grammar pattern, and prompted to do so, most pupils in the worked example classes were able to achieve this. However, further transfer was mixed. For example, teachers reported that it was a 'big step' for pupils to then use the patterns accurately when they were asked to incorporate the patterns into a piece of writing with multiple 'success criteria', or to writing activities or assessments in which pupils were not prompted to use the grammar pattern(s).

Some teachers in the worked example approaches described examples of 'far transfer', in which they reported that reviewing how a professional author or speaker had used a grammar pattern, and the opportunity to practice applying this in a low-stakes activity gave pupils the confidence to experiment with using different sentence structures during independent writing tasks and assessments. Teachers in the worked example approaches also reported that, after the sessions, some pupils were showing greater consideration to what they were writing, based on the effect and end goal they were aiming to achieve. These pupils were also viewed to be paying greater attention to how they were constructing their texts by considering what patterns would fit well with different parts of their writing when reviewing and editing their work. Teachers reported that this had made pupils' writing more thoughtful, more interesting, and better structured. These outcomes are aligned with the rationale for worked approaches, which aimed to develop not just use of a specific grammatical form, but also sensitivity to language choices. They were supported by having a teacher-made success criteria, which encouraged pupils to include different patterns within their writing.

> *It has been achievable for the students, when they are reflecting on their work, they had pride in their piece of writing, they know the quality has been improved because they have applied the structure, you see their sense of achievement, they can see the difference in their writing.* (Teacher, Responsive worked approach)

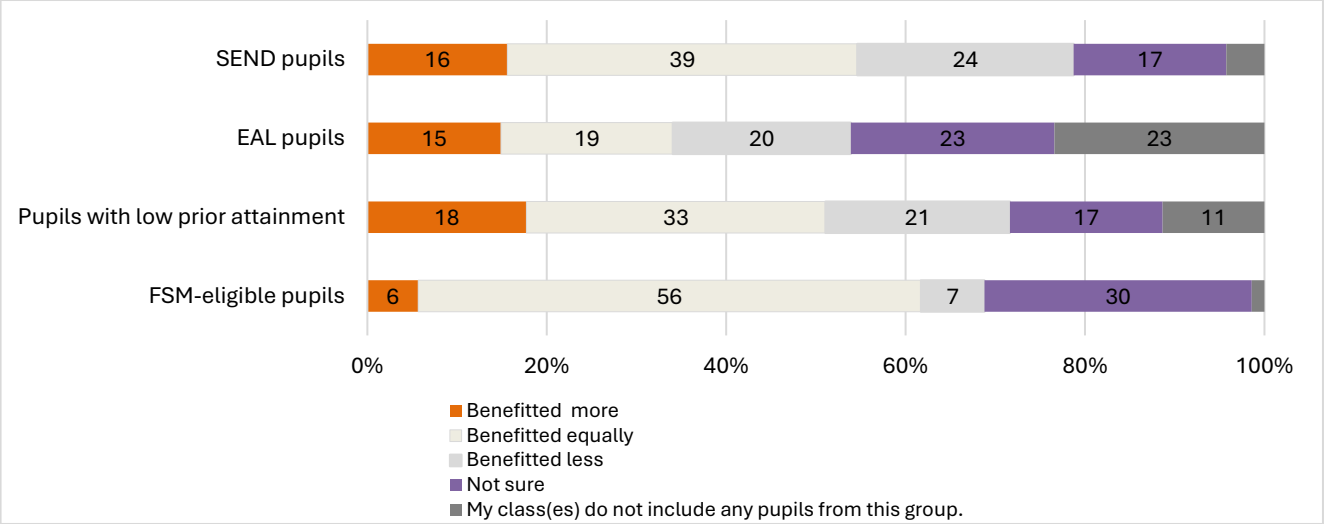
Teachers linked the skills that pupils were developing, such as the choice and effect of a grammar pattern, to developing both metacognition and the skills required for textual analysis in GCSE English literature.

Some teachers reported occasional examples of pupils transferring the patterns to different writing purposes (i.e. transferring the narrative fiction patterns beyond creative writing or persuasive speech examples beyond persuasive writing), particularly for higher attaining pupils, but these examples were limited.
Teachers highlighted the importance of revisiting the grammar patterns across different writing purposes and linking to lesson content so that they became embedded within pupils' writing. Teachers also reported that while they understood the guidance on not focusing on the names of the grammar patterns, labelling the patterns would support pupils' future recall and implementation.

> *I think it's really good that they're being exposed to lots of different things [grammar patterns], and being made aware that they can change their writing based on them. I'm not really convinced that they're going to be, without further input from us, a lot of further input, that they will be remembering to put this into their writing in extra units that we're doing later in the year.* (Teacher, Responsive worked approach)

To understand any variation in outcomes, the endpoint survey asked teachers to compare how FSM pupils, pupils with low prior attainment, EAL pupils, and SEND pupils, had benefited in comparison with their peers. For FSM pupils, pupils with low prior attainment, and SEND pupils, teachers most commonly reported that these pupils had benefited equally to other pupils (56% for FSM-eligible pupils, 33% for pupils with low prior attainment, and 39% for SEND pupils) with similar proportions reporting that these pupils benefited either more or less than their peers. For EAL pupils, a similar proportion of teachers reported that EAL pupils benefited equally (19%), less (20%), or more (15%) than their peers. Results were similar across the three approaches. Details are shown in Figure 14 below.

Figure 14: Comparative perceived impact for specific groups of pupils



Data from endpoint survey (N=141). Percentages may not add to 100% due to rounding.

Teachers in case study schools only commented on differential pupil outcomes based on attainment, however they perceived prior attainment to be a key moderator for pupils to engage in and benefit from the grammar approaches. There was consensus across teachers in case study schools that higher and middle attaining pupils had grasped the grammar patterns and transferred them into their writing more successfully than lower attaining pupils. They reported that while higher attaining pupils were using the grammar patterns consciously and purposefully to improve their writing, there was a sense that lower attaining pupils were 'shoehorning' them into the writing in order to fulfil the writing criteria, but without a clear understanding of why they were using the patterns.

To understand pupils' perceptions of their learning, the pupil focus groups undertaken with pupils in the systematic worked and responsive worked approaches[12] included an activity for pupils to create a 'wisdom wall' to display what they felt they had learned or improved on, from using grammar examples. Identified outcomes fell into three main themes: i) improved use of grammar patterns; ii) improvements to writing; and iii) wider outcomes, discussed below, which largely aligned with the intended outcomes cited in the trial logic model. These were generally similar across both the systematic worked and responsive worked approaches, though any differences are noted below.

Improved use of grammar patterns

Pupils reported improvements to their understanding, knowledge, and use of a range of grammar patterns. Pupils who received the systematic worked approach named specific patterns and aspects of grammar they had learned, including relative clauses, pre- and post-modifying nouns, and short sentences. Pupils felt confident knowing the writing contexts they could use a grammar pattern, and in using multiple grammar patterns in a piece of writing.

> *When we have to do descriptions, we always get a success criteria so it might say that we need to include five different devices [grammar patterns] so knowing more than one device helps because we can include more and compare things and describe things better than just using ordinary words.* (Pupil, Systematic worked approach)

Improvements to writing

Similarly, pupils from both worked example approaches identified that their writing had improved resulting from engaging with the grammar patterns. Pupils reported that they were using the grammar patterns to make their writing more creative (e.g. to create imagery and make writing more lively) and persuasive (e.g. through creating tension and making writing more powerful). Pupils said they considered the effect that their writing had on the reader. Pupils felt that they were better able to

[12] As discussed in the 'IPE' subsection in the 'Methods' section, due to unforeseen practical constraints during visits, we were not able to undertake focus groups with pupils from the non-worked approach.

structure their writing, which improved its fluency for a reader. The use of grammar patterns was also seen to make writing more interesting, which would support a readers' engagement with the text.

> *I think it makes our writing more detailed, because we know all these new patterns to use, it makes it more detailed, and you know I was talking about hooks, you can hook the reader onto it more*. (Pupil, Responsive worked approach)

Pupils reported a range of additional improvements to their writing and grammar skillset, including better use of metaphors and similes, improved spelling and punctuation, and expanded vocabulary. While these were not expected outcomes of the trial, pupils perceived the grammar sessions to have led to these improvements, which were in turn helping to improve pupils' writing.

Wider outcomes

Pupils identified wider outcomes of the grammar patterns, which suggested learning from the sessions was being transferred across tasks, and the curriculum. Pupils reported transferring the patterns to writing tasks in other subjects, which required descriptive or persuasive devices such as history, geography, and science.

> *I use a lot of the grammar in science when I'm writing examples and conclusions of experiments*. (Pupil, Systematic worked approach)

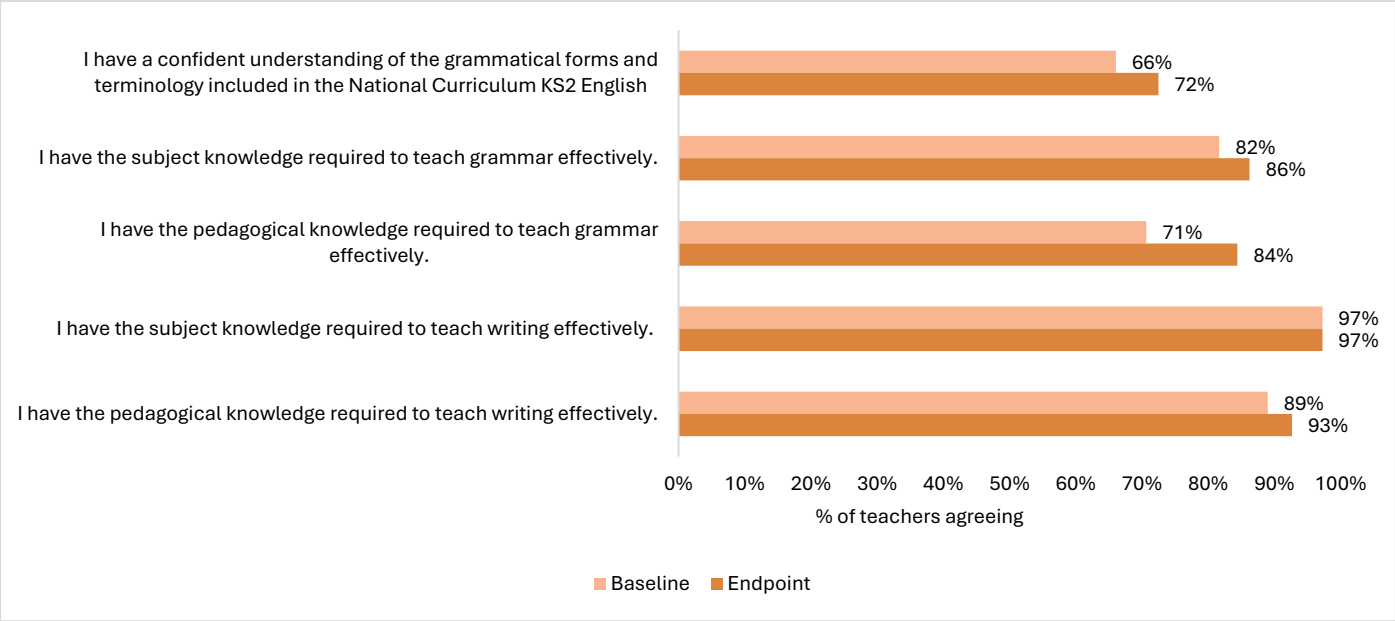
They were also using the grammar patterns to improve their writing in assessments. By working through the structure and features of the grammar patterns, pupils felt their text annotation skills had improved and they felt better able to recognise grammar patterns when reading texts (outside of the grammar sessions).

Perceived outcomes for teachers and schools

Knowledge for teaching grammar and writing

The baseline and endpoint surveys asked teachers a series of questions about their perceived knowledge and understanding for teaching grammar and writing. To understand any changes over the course of the trial we analysed matched responses from teachers who responded to both surveys (Figure 15). At both time points, agreement was highest with the statement related to subject knowledge required to teach writing effectively (97% strongly agreed/agreed at both time points). Agreement was lowest at both time points with the statement related to confidence in understanding of the relevant appendix in the Key Stage 2 English national curriculum (baseline 66%, endpoint 72%). Improvements were seen in relation to teachers' agreement that they had the pedagogical knowledge required to teach grammar effectively (71% baseline, 84% at endpoint). Teachers' perceptions were broadly similar across the allocated approaches.

Figure 15: Proportion of teachers who agreed with statements about their knowledge for teaching grammar and writing

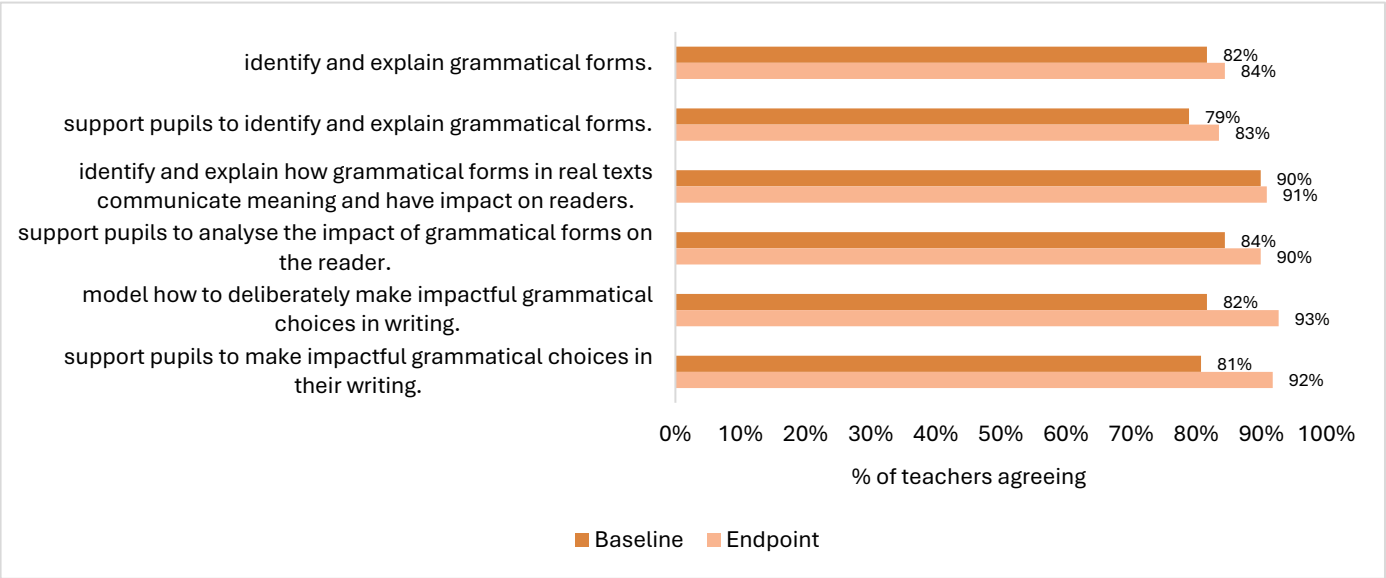


Matched teacher data from the baseline and endpoint surveys (N=109). To what extent do you agree or disagree with the following statements? Due to rounding, not all percentages sum to 100%.

Confidence teaching grammar

The baseline and endpoint surveys asked teachers to comment on their ability to teach grammar across six grammar pedagogical items. To understand any changes over the course of the trial we analysed matched responses from teachers who responded to both surveys (Figure 16). At baseline, a high proportion of teachers were confident in each of the abilities listed, as might be expected from the experience and specialism of trial teachers. At endpoint, a slightly higher proportion of teachers (c. 10%) agreed they were confident in modelling how to deliberately make impactful grammatical choices in writing (endpoint 93%, baseline 82%) and supporting pupils to make impactful grammatical choices in their writing (endpoint 92%, baseline 81%).

Figure 16: Proportion of teachers who agreed with statements about their confidence in teaching grammar



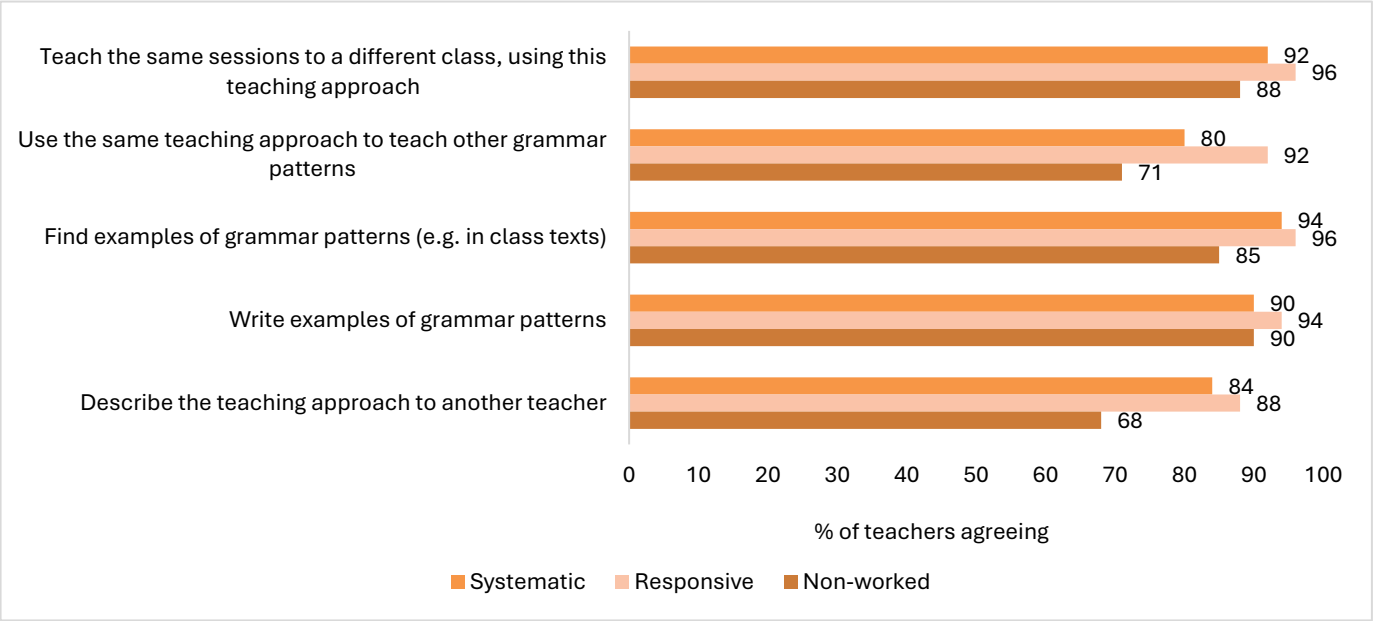
Matched teacher data from the baseline and endpoint surveys (N=109). Please indicate how strongly you agree or disagree with the following statements: 'I am confident in my ability to:' Due to rounding, not all percentages sum to 100%.

Confidence to use the grammar approaches in the future

The endpoint survey asked teachers to indicate the extent to which they agreed with a series of statements related to implementing grammar teaching, following the trial teaching approaches, in the future. Overall, over three-quarters of teachers agreed or strongly agreed they were confident in using elements of their allocated approach, however, there were

differences by grammar approach (Figure 17). Teachers in the worked approaches were more confident than those in the non-worked approach about using their approach to teach other grammar patterns, finding examples of grammar patterns, and describing their teaching approach to another teacher.

Figure 17: Teachers' confidence after completing the trial teaching



Data from the endpoint survey. 'I am confident in my ability to:' (Responsive worked N=49; systematic worked; N=50; non-worked N = 42).

There was limited data from teacher focus groups on the outcomes for teachers specifically as discussions tended to gravitate towards wider departmental outcomes described below.. However, teachers did report that they felt more informed about how to teach grammar explicitly and the short sessions through the trial had been an achievable way of including grammar teaching within lessons. Teachers also reported that they had been able to create clear success criteria (e.g. number of grammar patterns used, and correctly), both for when pupils peer reviewed and when teachers themselves marked creative and persuasive writing.

Over half (58%) of all teachers agreed or strongly agreed that they would continue to use their allocated teaching approach after the trial, though this proportion was higher for the worked example approaches (systematic worked 69%; responsive worked 62%; non-worked 40%). Similar trends could be seen in teachers' agreement (61%) that they would recommend their approach to another teacher (systematic worked 69%; responsive worked 68%; non-worked 46%).

Teachers involved with the focus groups shared their department's motivations for engaging with the trial. While explicit grammar teaching of grammar is not required by the Key Stage 3 curriculum, teachers identified a need for an approach to teach grammar in order to support pupils to become better, confident writers. Teachers hoped that participating in the trial would support either the development or replication of one of the approaches, and findings from the focus groups suggest that this had been achieved.

Case study schools said they would take forward elements of the systematic worked and/or responsive worked approaches. For example, some had plans to deliver one discrete lesson per week on grammar patterns, while others planned to incorporate grammar teaching into daily lessons, linked with the lesson content. None of the case study schools reported that they would be implementing the non-worked approach as delivered in the trial, because of the lack of opportunity for pupils to practice using patterns in their own writing. Teachers in case study schools reported that grammar teaching would be incorporated into schemes of learning across Key Stage 3, with the intention that by Key Stage 4, pupils would be confident using the patterns and creating their own style of writing, as well as being able to analyse the effect of a pattern in their own and other's writing. The aim being that this would enable them to access the higher grades at GCSE. Teachers reported that it would also be important for grammar to be taught across topics, not just creative and persuasive writing, so that pupils could see the use of grammar patterns for a range of writing purposes.

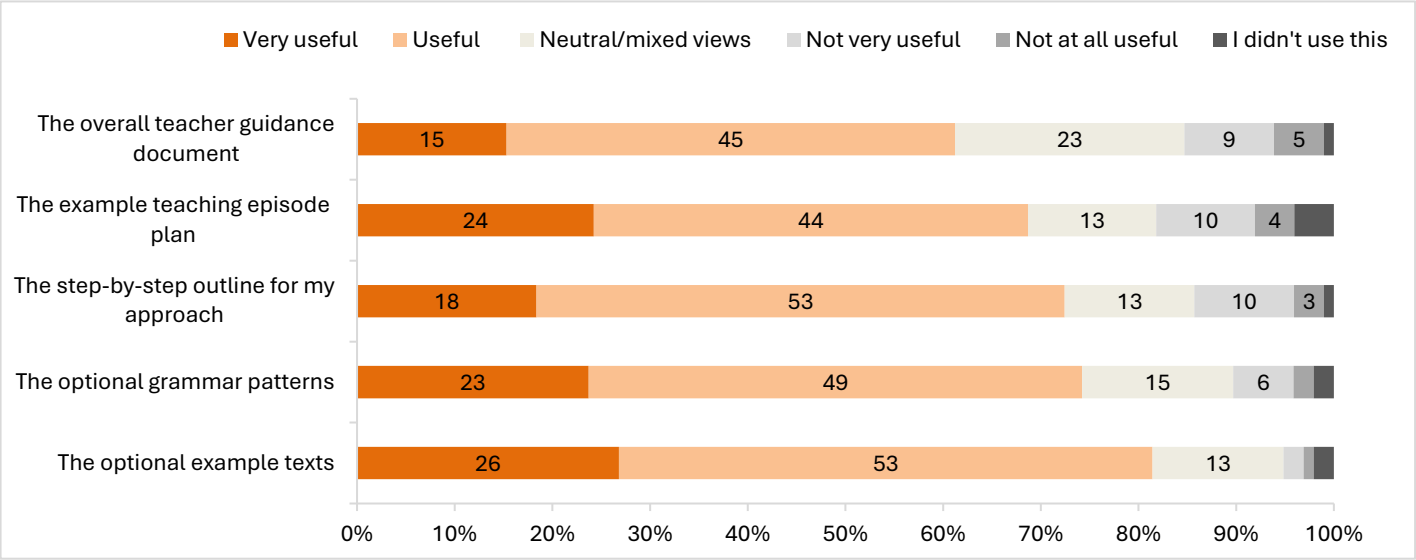
We've taken the lessons learned from it and then really tried to make it work. And I think I think that's probably, like it's the biggest testament I can give to it is that we are now adapting and making this part of our every day, every year curriculum. (Head of English, Systematic worked approach)

Trial design

Teacher use of guidance and support

To support their implementation, teachers involved in the trial received guidance and materials to support them in teaching grammar following their allocated approach. The endpoint survey asked teachers to indicate how useful they had found each of the materials they received (Figure 18). At least three-fifths of teachers perceived each of the materials to be useful. Overall, the optional example texts were perceived to be the most useful material (79%), followed by the optional grammar patterns (72%), the step-by-step outline for each approach (71%), and the example teaching episode in the teacher guidance was useful (68%). Three-fifths of all teachers had found the overall teacher guidance document for their approach useful (60%), but around a quarter (23%) had neutral/mixed views on this. Across each of the materials, small proportions indicated that they had found them 'not at all useful' or had not used them.

Figure 18: Teachers' views of elements of the teacher guidance



Data from the endpoint survey (N=144). Please rate each of the following teacher materials in terms of supporting you to teach your allocated approach. Due to rounding, not all percentages sum to 100%.

There were several differences in how useful teachers across the three grammar approaches found each of the guidance materials. Teachers allocated to the responsive and systematic worked approaches found the optional example texts (responsive worked 86%, systematic worked 84%) and example teaching episode plan (responsive worked 75%, systematic worked 72%) more useful than teachers allocated to the non-worked approach (72% and 59%, respectively). Teachers allocated to the non-worked approach were more likely to indicate that these materials were not useful, or that their views on these were neutral/mixed. Across the three approaches, teachers allocated to the responsive worked approach were more likely to indicate that they had found the optional grammar patterns useful (80%), followed by teachers allocated to the systematic worked (72%) and non-worked (69%) approaches.

Around three-fifths (61%) of all teachers agreed/strongly agreed that they received enough information/support to participate in the trial but around a quarter (24%) disagreed/strongly disagreed with this statement. There were no notable differences in levels of agreement by grammar approach.

Around three-fifths (61%) of all teachers also agreed/strongly agreed that they had been able to use the optional example texts (grammar pattern examples) without adapting them; however, just over a quarter (27%) disagreed/strongly disagreed with this statement. By grammar approach, teachers allocated to the responsive worked and systematic worked approaches of the trial were more likely to disagree (33% and 32%, respectively), thus indicating that they had needed to adapt the example texts, compared to teachers allocated to the non-worked approach (17% disagreed/strongly disagreed).

Teacher focus groups in case study schools provided further insights into their views on the guidance and materials provided to support their implementation of the grammar approaches. Teachers spoke positively of the authentic texts example bank. Teachers across all case study schools reported that they had used these examples rather than sourcing their own, which they said they would not have had time to do and felt that the provision of these examples had mitigated too much of a negative impact of the trial on workload. They praised the examples for exposing pupils to high-quality literature from well-renowned authors and influential speakers.

> *Having the bank of resources, that's really nice to have as a teacher, here is something special for them to emulate. Having that done for you is so good because often, half the battle is trying to find a good example. You've got some Terry Pratchett and Dickens and some really lovely, descriptive, thoughtful authors in there which was really nice.* (Teacher, Systematic worked approach)

Similarly, case study teachers found the guidance document, which contained dos and don'ts, a lesson rubric, and a sample lesson plan, clear and easy to follow, suggesting that case study teachers had more positive views than the broader group who responded to the survey. They reported that they required time to thoroughly review this guidance to ensure they fully understood how to deliver the grammar sessions in line with their allocated approach, but the detail provided in the guidance had supported interpretation and delivery with fidelity, for example, through ensuring that the focus on the session was on the grammar patterns, rather than the subject terminology. Teachers also appreciated the trust given to them to receive the approach guidance and use their professional judgement to teach the sessions following this, rather than receiving a Continuing Professional Development (CPD) course out of school.

Teachers fed back that several additions to the guidance and materials could have made it even more useful. Although teachers had been able to easily copy the authentic text examples into their own lesson PowerPoints, they would have liked to have received ready-made lesson plans and PowerPoints for each grammar pattern, which they could have used. Teachers of EAL pupils, lower-ability pupils/nurture groups reported that the authentic text examples were inaccessible for pupils working below expected Year 7 level (e.g. those accessing phonics) and would therefore, have appreciated simpler text examples and guidance on adapting lessons for these pupils.

Teachers in case study schools were asked what guidance they would need to continue using the grammar approaches in the future. Teachers were satisfied with the guidance, sample session plan, and rubric but identified the need for more authentic text examples. Teachers said it would be helpful to have text examples directly related to the texts and topics being studied in the curriculum, not only for Year 7 but also for other year groups (in line with plans to implement a chosen grammar approach across the department, for all year groups). They would also find it beneficial to have similar example banks for other topics.

> *Any more examples. Anything that makes resourcing something like this easier for professionals. Because in busy schools with lots of kids, we all know what it's like, if you want staff to buy in, you've got to keep it really simple. You've got to give them everything they need and then you've got to make it really easy for them to comply.* (Head of English, Systematic worked approach)

Implementing different choices within a department

Teachers working in departments where teachers had been assigned across two or three of the grammar approaches were asked about their experiences of implementing different teaching approaches within the same department. Almost half (48%) of these teachers agreed/strongly agreed that it had been helpful to try out more than one teaching approach in their department. Just over a quarter (27%) neither agreed nor disagreed, and just 13% disagreed/strongly disagreed. Teachers were also asked if they had found it difficult to try out more than one teaching approach in their department—around two-fifths indicated that it had not been difficult to do this (41% disagreed/strongly disagreed with the statement). In comparison, just over a quarter (27%) agreed/strongly agreed that it had been difficult to try out more than one teaching approach. Table 33 displays responses to these questions by an allocated grammar approach.

Table 33: Teachers' views on implementing more than one teaching approach, by allocated grammar approach

	Trying out more than one teaching approach has been helpful			It has been difficult to try out more than one teaching approach		
	Responsive worked	Systematic worked	Non-worked	Responsive worked	Systematic worked	Non-worked
Strongly agree / agree	54%	49%	39%	15%	35%	32%
Strongly disagree / disagree	7%	16%	14%	44%	37%	43%
Neither agree nor disagree	24%	23%	36%	29%	19%	18%
Don't know / want to say	15%	19%	11%	12%	9%	7%

Compared with the positive feedback in the scoping phase, teachers in case study schools had more mixed opinions over the experience of testing multiple approaches within the same school. In line with the survey findings, the main benefit was perceived to be that, at the end of the trial teachers in the same department would collectively have experienced each of the three approaches and would be able to compare and contrast experiences of implementation and outcomes and decide which approach to take forward. In comparison, the downside of testing multiple approaches was related to implementation during the trial. Teachers who reported that they would have liked randomisation to have been done at the department level felt this would have helped with workload, as the creation of resources would have been shared more widely across the department (discussed in 'Effect on teacher workload'). Teachers felt that not discussing across approaches went against their departmental ethos of conferring on new teaching practices, sharing experiences, and outcomes. Where multiple teachers within departments had been allocated to the same approach, teachers reported benefiting from discussing best practice and supporting one another with implementation. This was facilitated by the receipt of curriculum development or departmental time. Teachers who were the only ones allocated to their approach had found it particularly challenging and would have liked to have a colleague(s) to double-check terminology and their understanding of implementation.

Effect on teacher workload

The endpoint survey sought to understand the impact that participating in the trial had on teachers' workload. Just over two-fifths (43%) agreed/strongly agreed that participation in the trial significantly increased their lesson planning workload. Around a third (31%) indicated that participating in the trial had not increased planning workload. These proportions were comparable across each of the three grammar approaches.

There was consensus across interviewed teachers that trial participation had increased their workload, through planning and creating resources for this new, additional activity. Teachers of lower-ability and mixed-ability classes experienced additional workload in simplifying and scaffolding the session for their pupils. Teachers felt that the increased workload was most noticeable at the start of the trial, as they got to grips with understanding their allocated approach, the grammar patterns, and considered how to implement the grammar sessions. This process became smoother as they became more familiar with using the grammar approaches. To manage workload, they tended to report that teachers within the same approach had buddied up and took it in turns to plan the sessions, create the presentation slides, and any resources (however, they were also cases of the trial lead planning all the sessions, which they shared with teachers). Teachers would have liked the trial to come with all the sessions planned and resources prepared, however, were grateful for the sample session plan, which they found easy to emulate. They also reported that the example banks were invaluable and without these authentic text examples, they would have struggled to manage the additional workload of sourcing or creating their own.

Effect on other teaching

Findings from the endpoint survey suggest a considerable impact of the grammar sessions on the displacement of other lesson content, with almost three-quarters of all teachers agreeing/strongly agreeing that participating in the trial meant

having to drop or reduce content elsewhere. Just 15% disagreed that they had to do this. This appeared to affect teachers allocated to the systematic worked approach of the trial the most—86% reported agree/strongly agree, compared to 73% of non-worked teachers and 69% of responsive worked teachers.

Cost

As a Teacher Choices trial, a cost evaluation was not undertaken. If different arms were to be implemented in schools, likely costs would be teacher time to understand the approach, preparation to develop materials for pupils, and about five hours of teaching time (20 x 15-minute sessions), assuming teachers followed the guidance on session timing and sequencing.

Conclusion

Table 34: Key conclusions

Key conclusions	
1.	There was no evidence of meaningful differences between approaches to using examples on pupils' writing assessment scores.
2.	There was no evidence of meaningful differences between the approaches to using examples on the writing assessment scores of pupils eligible for free school meals (FSM).
3.	There was no evidence that prior attainment influenced the effect of different teaching approaches on pupils' writing assessment scores.
4.	The teaching approaches represented a substantial change to usual practice for many teachers. Teachers reported that a sustained focus on grammar patterns within text was new to their teaching and their classes, particularly elements of worked examples, such as modelling the step-by-step construction of a grammar pattern or asking pupils to follow that step-by-step construction in their writing. Given this substantial change, additional support for teachers may have been needed to achieve sufficient contrast between the approaches.
5.	Teachers in the worked example approaches perceived that most pupils could successfully use a grammar pattern in their writing when this was highly scaffolded. However, teachers perceived that pupils rarely transferred use of the taught grammar patterns into more general writing composition tasks and suggested pupils would need additional support to do so.

Impact evaluation and IPE integration

Evidence to support the logic model

Considering the implementation of each approach, the systematic worked examples approach was generally implemented by teachers as intended. However, the IPE indicated that the responsive worked approach was implemented more similarly to the systematic worked approach than intended. The aim was that teachers in the responsive worked approach would integrate the grammar sessions into the remainder of the lesson, rather than teaching it as a separate activity, which is what over three-quarters of this group reported doing 'all or most' of the time. As discussed in the 'Limitations and lessons learned' section below, it is likely that the short period of time from randomisation to the implementation period contributed to teachers (in all arms) making greater use of the example bank provided than expected. This is likely to have been a particular limitation for the responsive worked arm, as true integration of the use of worked examples would depend on finding examples from the texts they were studying or constructing examples relevant to the rest of the lesson. Finding and constructing their own examples would have been an unrealistic planning burden in the context of this relatively short notice change in teaching, especially as we also found that working with grammar patterns was new to many teachers.

The non-worked approach appeared to be difficult for teachers to implement in the context of writing. In contrast with the direct support for schema formation within the worked example approaches, where pupils were explicitly asked to construct text, which included the focus pattern, teachers in the non-worked group were asked not to include a writing exercise after teaching the example. However, two-fifths of teachers in the non-worked group reported doing this most or all of the time. The context of this may have influenced this as a writing-focused trial, and because teachers knew that there was a writing assessment at the end of the implementation period.

Considering pupil outcomes, although the impact analyses did not find any statistically significant differences in writing outcomes other than in one instance, in the FSM subgroup, inspection of the adjusted means indicates that there may be a pattern among the arms worth further exploration. There appears to be a trend across the writing analyses conducted, where the non-worked group consistently had the lowest adjusted mean score. We cannot be confident that this is a real difference, however, because many of the analyses take different slices of the same TSWA data, so more research would be required to determine this. It is also worth noting that the interaction with prior attainment hypothesis is more empirically justified than the FSM subgroup analysis (since prior evidence suggests modelling may work better for low attaining pupils) and this returned a null result. The significant result for the FSM subgroup is, hence, likely to be spurious.

There is some evidence from the IPE of very proximal outcomes in improved sophistication of grammatical constructions in pupils' focused writing within the trial teaching sessions. However, teachers reported that subsequent transfer of grammar patterns into writing tasks where use of the pattern did not immediately follow modelling, or was not explicitly prompted, was much rarer. Complementing this perception, the impact evaluation evidence indicates that there was no improvement in general writing composition from the use of worked examples, compared with the use of non-worked examples. There is also no evidence for the non-worked approach improving declarative knowledge of grammatical constructions, compared to the other arms. However, as the teacher's guidance for all three arms suggested focusing on the use and effect of the grammatical constructions, rather than grammatical terminology, a lack of effect here is not surprising.

Analysing examples of text and considering the effect on a reader is a common activity in the context of Key Stage 3 reading, and this appears to have carried through during the trial, with over three-fifths of teachers in all arms reporting this to be a feature of 'all or most' lessons. This is despite it not being a feature of the non-worked approach. However, it was expected to be a more common feature of the responsive worked approach than it actually was.

While pupils felt they had acquired better and more explicit understanding, teachers reported that it was rare for pupils to transfer the use of grammar patterns beyond the immediate focused writing context, into broader writing composition tasks. Teachers suggested that additional scaffolding and linking across sessions, for example, using retrieval practice for consolidation, would be needed to encourage this transfer. Similarly, the ten-week trial period did not allow time to build in 'backwards fading' to the working by gradually reducing the level of prompting, which may be necessary for pupils to use grammar patterns in broader contexts.

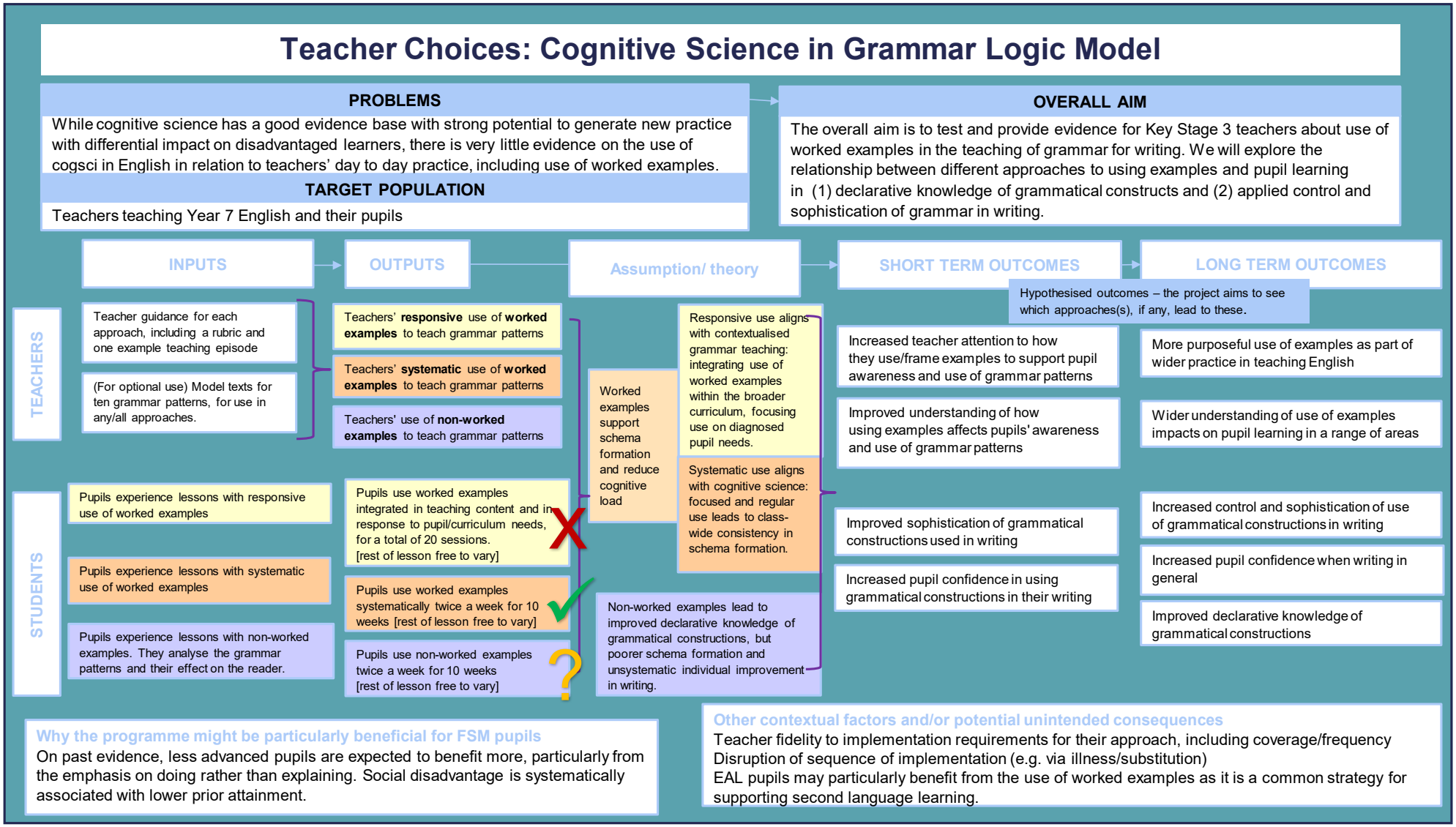
Considering the FSM subgroup, the initial analysis of writing scores for only FSM pupils showed a possible significant difference across the three teaching approaches. However, the post-hoc comparisons undertaken to directly compare each pair of teaching approaches showed only small differences in mean scores between the approaches (≤1 mark out of 40 maximum), which were not statistically significant. Therefore, we conclude that there is no evidence that the writing composition of FSM pupils differed after being taught with the different teaching approaches when measured with the writing assessment. Similarly, the interaction analysis showed no evidence that the approaches were working differently for FSM pupils and non-FSM pupils). There was no evidence that the FSM status of pupils changes the grammar assessment scores obtained under the different teaching approaches, or that the number of teaching sessions that FSM-eligible pupils were exposed to affected their writing scores.

Considering key moderators, the logic model anticipated that pupils with low prior attainment would benefit more from using worked examples, due to the reduced cognitive load associated with basic step guidance. In contrast, teachers reported that pupils with high prior attainment were better able to access the grammar patterns and use them in their own writing. This appears to be a context-specific effect, with teachers reporting that lower attaining pupils lacked the necessary prior knowledge to identify some grammar patterns (e.g. identifying nouns and adjectives). This may have created a very high cognitive load for initial engagement with the grammar patterns, explaining the discrepancy with the logic model.

In terms of teacher outcomes, a key outcome noted in the IPE is a substantial change in practice to focus on grammar patterns within text, and to use worked examples in this context, modelling step-by-step construction and encouraging pupils to include the pattern in writing. This is different from teachers reported usual practice, and most teachers (especially in the worked examples approaches) intended to continue using their approach in the future. While we cannot be sure due to the lack of a comparison group who did not use examples for teaching grammar, this overall change in teacher practice may have dwarfed the smaller differences between the tested approaches, especially considering the reduced contrast in the teaching approaches as implemented.

Overall, teachers responded positively to the example approaches, and felt they met a need to develop grammar teaching in Key Stage 3 English. They reported gains in knowledge for teaching grammar, and case study schools intended to embed grammar worked examples in their future teaching.

Figure 19: Annotated logic model (post-trial)



Interpretation

The evaluation showed no differential effect on pupils' writing composition from the different approaches to using examples. The confidence in this finding is limited by the high percentage of missing data (29% of all randomised pupils). Challenges in implementation, as well as challenges in transfer of learning, may have contributed to this result.

Evidence about the implementation of the approaches shows a reduced contrast between the teaching approaches, and uncertainty in whether teachers could teach as many sessions as intended. The reduced contrast between the approaches substantially limited the difference between the teaching experienced by pupils. Furthermore, as we are missing data for about half of pupils on the number of taught sessions, we cannot be sure how much of the approach was experienced by pupils, or whether this varied by approach. The evaluation therefore, did not provide any conclusive answers about the relative effectiveness of the approaches compared to each other.

Furthermore, teachers indicated that all of the approaches involved a more extensive and sustained focus on grammar patterns, compared with their usual teaching. This broader change in practice may have eclipsed the smaller differences between the approaches. However, in the absence of a comparison group who were asked to continue with their usual practice, we are unable to conclude whether *any* use of grammar pattern examples is more or less effective than teachers' usual practice: the approaches could have had a positive effect, negative effect, or no impact. Despite this, the IPE indicated promise for these approaches, with teachers open to continuing to use worked examples in the future. In light of the findings, we recommend that further research be conducted into the use of cognitive science approaches in the classroom. The approaches were well received by teachers and therefore, evidence about how these approaches compare to the usual practice of not working with grammar pattern examples, suggested by teachers participating in the trial, would benefit the teaching community.

Aside from the idea of cognitive science approaches, it was interesting to see that teachers and pupils were receptive to increased explicit teaching of grammar in Key Stage 3. Grammar is a key focus at Key Stage 2, but less so from Key Stage 3 onwards and teachers suggested that building grammar into Year 7 lessons was a useful and familiar bridge to support transition into English lessons at secondary school.

There is also a wider question about the focus and approach of Teacher Choices trials to consider. Teacher Choices trials were originally conceived to test common classroom practices against each other, rather than in comparison to current practice. However, as emerged in the IPE, none of the approaches here were common, and all were considered novel to some extent by participants. While we expected that some use of examples would be new to at least some teachers, the extent of this meant that the trial needed to supply guidance for each of the arms to the teachers.

Teachers did not adapt the approaches and content taught as much as we expected, which appeared to be due to a combination of time/workload pressures and a perception that they needed to follow the guidance exactly. Indeed, reflecting on the feedback and implementation of the trial, more support (see 'Limitations and lessons learned' section below) and perhaps a 'learning period' to acclimatise to the use of new approaches may scaffold implementation prior to the trial period. For example, a common concern from teachers in this trial was the extent to which they could adapt the approaches and the materials. If Teacher Choices trials aim to support a complex change in practice, we consider that more time to digest the allocated approach and additional avenues throughout the trial to access reassurance and advice from the research team could be beneficial. Alternatively, Teacher Choices trials need to focus on choices, which are simple to implement well.

Limitations and lessons learned

The trial was conducted in real-world conditions by schools and teachers who volunteered to sign up for the project. We expect that this group of teachers were more positively disposed towards changes in practice, and to developing their grammar teaching, than the general population of teachers might be.

The scoping phase aimed to explore the best subject (English, maths, or science) and phase (Key Stage 2 or Key Stage 3) for exploring worked examples. This means that the Study Advisory Board and the evaluation team could not include experts

from across these subjects and phases. While secondary English specialists were added to the Study Advisory Board and research team during trial set-up (and after scoping), an earlier dialogue would have supported an intervention design, which integrated more clearly into secondary English.

This Teacher Choices trial entailed a high level of novel practice. In choosing to focus on grammar in Key Stage 3 English (where grammar is not explicitly emphasised in the curriculum, and most English teachers are specialists in literature rather than language), we significantly extended the focus on grammar patterns for writing for trial teachers and classes, across all three approaches. Similarly, in choosing to extend the research on worked examples beyond maths and science into English, we were asking teachers to translate the elements of worked examples for use in a different subject and pedagogical tradition.

The scoping phase suggested that English teachers made extensive use of examples in their teaching that grammar was a relatively low priority for Key Stage 3 English teachers, and that grammar was most commonly taught either in separate activities/lessons, or integrated into lessons. Based on this, we decided to proceed with a trial comparing different approaches to using examples in grammar teaching, rather than comparing the use of examples with the non-use of examples. However, the scoping phase did not explore the nuances of how Key Stage 3 English teachers used examples in the specific context of grammar teaching. Similarly, as the three approaches to using examples were designed at the end of the scoping phase, we did not explore teachers' familiarity with or prior use of these specific approaches, or the use of grammar patterns, until the trial baseline survey. The data collected indicated that worked approaches to using grammar patterns were new to most English teachers in the trial. Similarly, through case study visits we learned that teachers found it difficult to use non-worked examples in the context of teaching writing, as it precluded pupils rehearsing writing themselves. While English teachers commonly used the features of worked examples, such as modelling a step-by-step construction and asking pupils to follow these in text construction, the use of grammar patterns was novel. While the teacher guidance was shared with practitioners during trial set-up for their feedback, to check clarity and feasibility of implementation, the agreed trial timeline precluded more extensive piloting. In future Teacher Choices trials, we would recommend a two-stage scoping phase, to allow for additional scoping activity around the specific parameters being proposed for evaluation (in this case understanding common practice in the use of examples in teaching grammar for writing in Key Stage 3).The decision to focus on a choice, which significantly extended practice had implications for interpreting the findings and for the provision of teacher guidance. As outlined above, without a comparison group who did not use grammar pattern examples, it is unclear whether the different approaches were equally effective or whether none of them had an impact.

Guidance from the EEF indicates the Teacher Choices trials should provide minimal teacher guidance in the form of written materials. However, interpreting how to use worked examples in the context of teaching English was a significant translation challenge for teachers in this trial. We tried to mitigate this by providing additional modelling for each approach, in the form of an example session plan. Further, we provided optional curriculum materials in the form of an example bank containing eighty examples for ten grammar patterns. In the absence of these resources, finding or creating examples would have entailed a substantial burden for teacher planning, in what we note was quite a short period from randomisation (and therefore, notification of allocated approach) to the start of the implementation period. While these examples were optional, almost all teachers reported using them. This limits the generalisability of the results, as teachers used one specific set of grammar patterns and examples, rather than different grammar patterns or examples tailored to the class and curriculum. Teachers also reported that they were limited in their ability to adapt their teaching of the supplied grammar patterns to the needs of their class, often because they were covering other texts in class. However, teachers reported that the resource bank was necessary to support their planning in a short time frame and to support their ability to model the use of grammar patterns.

A consideration for future Teacher Choices trials is whether to allow more dialogic engagement with teachers, for example, through an initial webinar, directly after teachers were given their allocated approach, and/or after a brief learning period to try out the materials. This could reduce the extent of written materials needed and engage teachers more directly with the rationale and distinction between the choices. In this trial, this could have been used as an opportunity to reassure teachers that they are encouraged to use their professional judgement in adapting their teaching for their class and curriculum context, minimising teacher concerns in this trial about 'getting it right' by closely following the optional aspects of the guidance (e.g. sample plan and example bank). We recommend that, where a Teacher Choices trial does involve

pedagogical activities that could be less familiar to the teachers involved, dialogic engagement and support with the activities is essential. This would be expected to increase the contrast between choices.

As we chose to randomise teacher-class units based on the feedback from the scoping phase, different teaching approaches were being tested within the same school. This design can potentially result in more contamination than a school randomised design, if teachers in the same school discuss their practice and share resources. To minimise contamination, participating teachers only received detailed guidance for their own allocated approach and were asked to avoid discussing the approaches with teachers allocated to other approaches until after the trial. Evidence from the IPE suggested that teachers had cooperated with this request, with 78% of teachers reporting that they had used their allocated approach (as they understood it) in all their sessions. Furthermore, during the interviews teachers reported that they had avoided any discussion of the different approaches. More broadly, Teacher Choices trials inherently rely on collaboration from teachers, who voluntarily choose to adopt or avoid particular pedagogies, which are broadly available for them to use, in order to contribute to a trial, rather than relying on external training or resources, as is commonly the case for programme trials, where access to the intervention can be more actively restricted. Overall, on this trial, we do not believe that there was any significant contamination and that reasons for the reduced contrast between arms lay elsewhere, as discussed. However, we acknowledge that, as adherence data was self-reported, we cannot rule out contamination between the approaches as a reason for reduced contrast. As each teacher was allocated to an 'active' teaching approach, we do not expect any experimental effects associated with allocation to a control group.

Considering the logistics of evaluation, we recommend that similar future evaluations streamline the number and burden of data collection activities to be commensurate with shorter Teacher Choices trials. Our endpoint data collection took place very near the end of the Summer Term, with teachers asked to complete endpoint pupil assessments, return session delivery logs, and complete the Teacher Endpoint survey in the same three-week period. In our reminders, we prioritised the pupil assessments (as the primary outcome measure) where these had not been completed. This meant that only about half of the teachers completed the session delivery logs (which provided the dosage and compliance data) and the endpoint teacher survey. While all schools that were expected to complete endpoint assessments returned at least some, there was substantial attrition due to whole classes not completing assessments (196 pupils, 2%), and pupil-level absence (655 pupils, 7%) on the day of the assessments. We expect this was partly due to timetabling changes (e.g. school trips and enrichment activities) at the end of the Summer Term. This contributed to an overall attrition rate of 29% for the primary analysis of the effect on writing composition. As missing data on the writing assessment was associated with Key Stage 2 attainment (Key Stage 2 GPS score), FSM, and EAL, we assumed that outcome data were MAR, and ran a sensitivity check for the primary analysis using imputed data. This sensitivity check had similar results to the primary analysis, and also showed a null effect on writing composition, suggesting that the primary finding is robust to the effects of attrition and missing data. A key impact of low response rates on the evaluation was that dosage data were missing for about half of the trial pupils. We therefore, believed that any compliance analysis would be non-robust and challenging to interpret. After consultation with the EEF and a Study Advisory Board member, we agreed not to run a compliance analysis. The pre-specified dosage analysis is reported in the impact evaluation results, though it is similarly limited by the missing dosage data. Providing a financial incentive for providing dosage data is likely to support a higher response rate. As the primary research analyses are based on an ITT model, these are not affected by the limitations in dosage/compliance data. However, we cannot draw conclusions about whether the observed null effects are due to unobserved factors in the usage/dose of the teaching approaches, for example, differential usage/dose between the three approaches, or low/no use across the approaches.

Considering the bespoke outcome measures, the markers for both assessments were blinded to treatment allocation, in line with recommended practice. Psychometric analysis of the primary outcome measure (TSWA), considering strand internal consistency, marker reliability, and frequency distribution showed that the functioning of the measure was psychometrically adequate. Psychometric analysis of the secondary outcome measure (NPGA), including item-internal consistency and frequency distribution showed that the functioning of the measure was psychometrically adequate. While a small proportion (0.73% of respondents) scored 30 out of 30 marks for the NPGA, this was below the 5% threshold recommended in the literature (Fisher Jr., 2007) for a ceiling effect, which affects measurement properties.

Future research and publications

Across the EEF writing trials (e.g. Torgerson *et al.,* 2018; Anders *et al.,* 2021), it has been challenging to find pedagogical approaches with a measurable impact on pupil writing composition. In particular, the transfer of learning over time and across writing contexts appears to be a common point of difficulty. Future research could explore current practice or potential approaches to encouraging transfer, including those based on strategies from cognitive science.

There is continuing interest in using cognitive science approaches across the curriculum. For example, the second priority (of 15) for teachers to gain more from cognitive science research, according to the recent survey by the Chartered College of Teaching was: 'How can cognitive science strategies support the retrieval and application of complex information, for example in literature or history teaching?' (Müller and Cook, 2023). Similarly, the EEF's practice review of writing (Grima *et al.,* 2024) identified the investigation of 'step-by-step' approaches to scaffold and model good writing as a priority for future research.

The scoping phase of the current project suggested the potential to move beyond grammar constructions to argument construction in essay writing. The cognitive science literature points to worked examples and schema formation as methods to establish basic templates of argumentation in higher level writing. This translation would have immense potential if it could support the development of advanced skills for young people studying in the humanities. However, to the extent that such extension of cognitive science beyond the familiar territory of worked examples in maths and science constitutes novel methods for teachers, they would require comparison to non-use of examples to demonstrate their value.

While programme trials focused on teacher change usually incorporate formal professional development, action planning, and ongoing dialogue with an external developer, Teacher Choices trials rely on brief written guidance as the primary impetus for a complex change in practice. Future Teacher Choices trials should consider incorporating mechanisms to support behaviour change in teachers, such as motivating goal-directed behaviour, teaching techniques, or encouraging embedding of practice (Sims *et al.,* 2021).

References

Anders, J., Shure, N., Wyse, D., Barnard, M., Frerichs, J. and Bohling, K. (2021) *'The Craft of Writing'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/projects/Craft_of_Writing_Evaluation_Report_Final.pdf?v=1743510829 (accessed 01 April 2025).

Atkinson, R.K., Renkl, A. and Merrill, M.M. (2003) 'Transitioning from Studying Examples to Solving Problems: Effects of Self-Explanation Prompts and Fading Worked-Out Steps'. *Journal of Educational Psychology*, 95: 4, 774–83. https://doi.org/10.1037/0022-0663.95.4.774

Barbieri, C.A., Miller-Cotto, D., Clerjuste, S.N. and Chawla, K. (2023) 'A Meta-Analysis of the Worked Examples Effect on Mathematics Performance'. *Educational Psychology Review*, 35: 1, 11. https://doi.org/10.1007/s10648-023-09745-1

Bartlett, F.C. (1932) *'Remembering: A study in Experimental and Social Psychology'*. Cambridge: Cambridge University Press.

Blake, J. and Shortis, T. (2010) 'The Readiness is All the Degree Level Qualifications and Preparedness of Initial Teacher Trainees in English'. *English in Education*, 44: 2, 89–109. https://doi.org/10.1111/j.1754-8845.2010.01068.x

Braun, V. and Clarke, V. (2006) 'Using Thematic Analysis in Psychology'. *Qualitative Research in Psychology*, 3: 2, 77–101. https://doi.org/10.1191/1478088706qp063oa

Chen, H. and Myhill, D. (2016) 'Children Talking About Writing: Investigating Metalinguistic Understanding'. *Linguistics and Education*, 35, 100–8. https://doi.org/10.1016/j.linged.2016.07.004

Chen, O., Kalyuga, S. and Sweller, J. (2015) 'The Worked Example Effect, The Generation Effect, and Element Interactivity'. *Journal of Educational Psychology*, 107: 3, 689–704. https://doi.org/10.1037/edu0000018

Data Protection Act 2018, c.12. Available at: https://www.legislation.gov.uk/ukpga/2018/12/contents (accessed 26 September 2025).

Demack, S. (2019) *'Does the Classroom Level Matter in the Design of Educational Trials? A Theoretical & Empirical Review'*. EEF Research paper No. 003. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/methodological-research-and-innovations/Does_the_classroom_level_matter.pdf?v=1706475091 (accessed 26 February 2024).

Department for Education (DfE). (2013) *'English Programmes of Study: Key Stage 3. National Curriculum in England'*. London: DfE. Available at: https://assets.publishing.service.gov.uk/media/5a7b8761ed915d4147620f6b/SECONDARY_national_curriculum_-_English2.pdf (accessed 26 February 2024).

Education Endowment Foundation (EEF). (2022) *'Statistical Analysis Guidance for EEF Evaluations'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1705995156 (accessed 26 September 2025).

Fisher Jr., W.P. (2007) 'Rating Scale Instrument Quality Criteria'. *Rasch Measurement Transactions*, 21: 1095. http://www.rasch.org/rmt/rmt211a.htm

General Data Protection Regulation (GDPR). (2016) *'Council Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (United Kingdom General Data Protection Regulation) (Text with EEA relevance)'*. Available at: https://www.legislation.gov.uk/eur/2016/679 (accessed 26 September 2025).

Graham, S., Bruch, J., Fitzgerald, J., Friedrich, L., Furgeson, J., Greene, K., Kim, J., Lyskawa, J., Olson, C.B. and Smither Wulsin, C.(2016) *'Teaching Secondary Students to Write Effectively (NCEE 2017-4002)'*. Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Available at: http://whatworks.ed.gov (accessed 26 February 2024).

Graham, S., Harris, K. and Chambers, A. (2016) 'Evidence-Based Practice and Writing Instruction: A Review of Reviews'. In: C.A. MacArthur, S. Graham, and J. Fitzgerald (eds.) *Handbook of Writing Research*. New York, NY: Guilford Press. pp. 211–26.

Grima, G., Hooper, A., Mullins, L., Redmond, B., Sharott, P. and Ashworth, B. (2024) *'Understanding Current Practice and Research Priorities in Teaching Writing: Practice Review'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/understanding_current_practice_and_research_priorities_in_teaching_writing.pdf?v=1743511219 (accessed 01 April 2025).

Jones, S., Myhill, D. and Bailey, T. (2013) 'Grammar for Writing? An Investigation of the Effects of Contextualised Grammar Teaching on Students' Writing'. *Reading and Writing*, 26: 8, 1241–63. https://doi.org/10.1007/s11145-012-9416-1

Kellogg, R.T. (1999) *'The Psychology of Writing'*. New York, NY: Oxford University Press.

Keselman, H.J. and Rogan, J.C. (1977) 'The Tukey Multiple Comparison Test: 1953–1976'. *Psychological Bulletin*, 84: 5, 1050–56. https://doi.org/10.1037/0033-2909.84.5.1050

Kyun, S., Kalyuga, S. and Sweller, J. (2013) 'The Effect of Worked Examples When Learning to Write Essays in English Literature'. *The Journal of Experimental Education*, 81: 3, 385–408. https://doi.org/10.1080/00220973.2012.727884

McLaren, B.M., van Gog, T., Ganoe, C., Karabinos, M. and Yaron, D. (2016) 'The Efficiency of Worked Examples Compared to Erroneous Examples, Tutored Problem Solving, and Problem Solving in Computer-Based Learning Environments'. *Computers in Human Behavior*, 55: Part A, 87–99. https://doi.org/10.1016/j.chb.2015.08.038

McLoughlin, N. (2008) 'Chapter 8. Creating An Integrated Model for Teaching Creative Writing: One Approach'. In: G. Harper and J. Kroll (eds.) *Creative Writing Studies: Practice, Research and Pedagogy*. Bristol: Blue Ridge Summit, Multilingual Matters. pp. 88–100. https://doi.org/10.21832/9781847690210-010

Müller, L-M. and Cook, V. (2023) *'Cognitive Science in Education: Teachers' Priorities for Research'*. London: Chartered College of Teaching. Available at: https://chartered.college/wp-content/uploads/2023/06/MullerCook_2023_FINAL.pdf (accessed 01 March 2024).

Myhill, D., Jones, S. and Lines, H. (2018) 'Supporting Less Proficient Writers Through Linguistically Aware Teaching'. *Language and Education*, 32: 4, 333–49. https://doi.org/10.1080/09500782.2018.1438468

Myhill, D., Lines, H. and Jones, S.M. (2018) 'Texts That Teach: Examining the Efficacy of Using Texts as Models'. *L1-Educational Studies in Language and Literature*, 18: 2, 1–24. https://doi.org/10.17239/L1ESLL-2018.18.03.07

Myhill, D. and Watson, A. (2014) 'The Role of Grammar in the Writing Curriculum: A Review of the Literature'. *Child Language Teaching and Therapy*, 30: 1, 41–62. https://doi.org/10.1177/0265659013514070

Newman, R. and Watson, A. (2020) 'Shaping Spaces: Teachers' Orchestration of Metatalk About Written Text'. *Linguistics and Education*, 60, 100860. https://doi.org/10.1016/j.linged.2020.100860

Ofsted (Office for Standards in Education, Children's Services and Skills). (2022) *'Curriculum Research Review Series: English'*. GOV.UK. Available at: www.gov.uk/government/publications/curriculum-research-review-series-english/curriculum-research-review-series-english (accessed 01 March 2024).

Perry, T., Lea, R., Rübner Jørgensen, C., Cordingley, P., Shapiro, K., Youdell, D., Harrington, J., Fancourt, A., Crisp, P., Gamble, N. and Pomareda, C.(2021) *'Cognitive Science in the Classroom'*. London: Education Endowment Foundation. Available at: https://potentialplusuk.org/wp-content/uploads/2022/02/Cognitive_Science_in_the_classroom_-_Evidence_and_practice_.pdf (accessed 01 May 2025).

Reiss, K.M., Heinze, A., Renkl, A. and Groß, C. (2008) 'Reasoning and Proof in Geometry: Effects of a Learning Environment Based on Heuristic Worked-Out Examples'. *ZDM*, 40: 3, 455–67. https://doi.org/10.1007/s11858-008-0105-0

Richey, J.E. and Nokes-Malach, T.J. (2013) 'How Much Is Too Much? Learning and Motivation Effects of Adding Instructional Explanations to Worked Examples'. *Learning and Instruction*, 25, 104–24. https://doi.org/10.1016/j.learninstruc.2012.11.006

Sims, S., Fletcher-Wood, H., O'Mara-Eves, A., Cottingham, S., Stansfield, C., Van Herwegen, J. and Anders J. (2021) *'What Are the Characteristics of Effective Teacher Professional Development? A Systematic Review and Meta-Analysis'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/pages/Teacher-professional-development.pdf?v=1727862311 (accessed 21 October 2024).

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Einbeck, J., Higgins, S., Culliney, M. and Demack, S. (2023) *'Improving Power Calculations in Educational Trials'*. London: Education Endowment Foundation. Available at: https://educationendowmentfoundation.org.uk/projects-and-evaluation/evaluation/eef-evaluation-reports-and-research-papers/methodological-research-and-innovations/improving-power-calculations-in-educational-trials (accessed 02 October 2023).

Smith, A., Aston, K., Schwendel, G., Ager, R., Brill, F., Thomas, M., Tolmie, A., Watson, A. and Poet, H. (2024) *'Cognitive Science Teacher Choices: Using Examples to Teach Grammar to Year 7 (A Randomised Controlled Trial). Evaluation Study Plan'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/eef_cognitive_science_teacher_choices_-_study_plan_-_may_2024_update.pdf?v=1758557631 (accessed 26 September 2025).

Sweller, J. (1994) 'Cognitive Load Theory, Learning Difficulty, and Instructional Design'. *Learning and Instruction*, 4: 4, 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Tarmizi, R.A. and Sweller, J. (1988) 'Guidance During Mathematical Problem Solving'. *Journal of Educational Psychology*, 80: 4, 424–36. https://doi.org/10.1037/0022-0663.80.4.424

Thurston, A., Topping, K.J., Tolmie, A., Christie, D., Karagiannidou, E. and Murray, P. (2009) 'Cooperative Learning in Science: Follow-up from Primary to High School'. *International Journal of Science Education,* 32: 4, 501–22. https://doi.org/10.1080/09500690902721673

Torgerson, C.J., Ainsworth, H., Bell, K., Elliott, L., Fountain, I., Gascoine, L., Hewitt, C.E., Kasim, A., Kokotsaki, D. and Torgerso, D.J. (2018) *'Calderdale Excellence Partnership: IPEELL. Evaluation Report and Executive Summary'*. London: Education Endowment Foundation. Available at: https://d2tic4wvo1iusb.cloudfront.net/production/documents/projects/IPEELL_1.pdf?v=1743508239 (accessed 01 April 2025).

Van Gog, T., Paas, F. and Van Merriënboer, J.J.G. (2006) 'Effects of Process-Oriented Worked Examples on Troubleshooting Transfer Performance'. *Learning and Instruction*, 16: 2, 154–64. https://doi.org/10.1016/j.learninstruc.2006.02.003

Watson, A.M., Newman, R.M.C. and Morgan, S.D. (2021) 'Metatalk and Metalinguistic Knowledge: The Interplay of Procedural and Declarative Knowledge in the Classroom Discourse of First-Language Grammar Teaching'. *Language Awareness*, 30: 3, 257–75. https://doi.org/10.1080/09658416.2021.1905655

Whitney, P. (2001) 'Schemas, Frames, and Scripts in Cognitive Psychology'. *International Encyclopedia of the Social & Behavioral Sciences*, 13522–6. https://doi.org/10.1016/B0-08-043076-7/01491-1

Wolf, B. and Harbatkin, E. (2023) 'Making Sense of Effect Sizes: Systematic Differences in Intervention Effect Sizes by Outcome Measure Type.' *Journal of Research on Educational Effectiveness,* 16: 1, 134–61. https://doi.org/10.1080/19345747.2022.2071364

Appendix A: Security classification of trial findings

Note: The security of this trial's findings is based on the combined peer reviewer summary assessment below. This assessment is an interim approach for this evaluation while the EEF continues to develop its approach to reporting evidence security for EEF Teacher Choices evaluations.

OUTCOME: Text-type Specific Writing Assessment (TSWA), a bespoke assessment of writing composition developed by NFER

Domain	Comments
Study design	The three-arm cluster randomised controlled trial at the teacher-class unit level was rigorously implemented to compare the impact of relevant teacher choices, with appropriate power calculations, randomisation, and allocation, all clearly reported alongside baseline balance. The choice of comparators is well considered and justified.
Attrition	Attrition for the primary outcome was moderate across the trial arms. These were relatively balanced across different arms (26%; 31%; 31%). Reasons provided are clear, and stem from withdrawal from the evaluation or non-completion of the final assessment. Missingness has been fully explored and accounted for, with analyses indicating missingness was at random.
Compliance & choice adherence	Teachers received detailed information on how to implement the teaching approach and adherence to their allocated choice was good, as a high number reported they had used their approach in each session (78%). Compliance was well defined but high levels of missing log data due to data collection difficulties (48.8%) limited its analysis and understanding on whether compliance resulted in higher impact. Contamination among approaches was investigated and reported. While some issues were identified, contamination was not determined to be a significant factor which eroded the choice contrast.
Primary outcome and effects	The primary outcome was appropriate for the evaluation design and was well reported. Appropriate reporting of missing data, robustness, and further analyses was undertaken. The measure was constructed from validated and reliable measures. However, it was not possible to validate the measure prior to the trial. Psychometric analysis shows internal adequacy of measure.
Contextual factors	Contextual factors were well captured by the IPE and discussed in the report. Adjusted analyses presented as appropriate. Appropriate analysis and discussion of observed small ceiling effects in the secondary outcome measure undertaken.
Transparency in reporting	The study was pre-registered and there was adherence to the pre-specified Statistical Analysis Plan. Minor deviations have been well reported and justified.

Overall assessment

Overall, this Teacher Choices trial was carefully designed and carried out, offering robust findings about the impact of the different teaching approaches. The analyses were consistent with the pre-registered study plan and deviations clearly explained and reported. The study involved a fair comparison between three teaching approaches and results were well reported and supported, the threats are clearly noted, and conclusions drawn are grounded in the analyses presented. While missing data for the primary outcome averaged 29%, which is considered moderate, it was evenly distributed across the three teaching approaches. This balance reduces the risk of bias and supports the reliability of the results. Despite the possibility of contamination due to randomisation at the teacher level, contamination was not determined to be a significant factor which eroded the contrast between choices.

Appendix B: Frequency distributions for outcome measures

Table B1: Frequency distribution for the writing assessment (TSWA)

Score	Frequency	Percentage	Cumulative percentage
0	9	0.14	0.14
1	4	0.06	0.20
2	15	0.23	0.43
3	35	0.53	0.96
4	32	0.49	1.45
5	60	0.91	2.36
6	74	1.13	3.49
7	108	1.64	5.13
8	130	1.98	7.11
9	141	2.15	9.25
10	177	2.69	11.95
11	198	3.01	14.96
12	239	3.64	18.60
13	305	4.64	23.24
14	323	4.92	28.16
15	361	5.49	33.65
16	371	5.65	39.30
17	408	6.21	45.51
18	429	6.53	52.04
19	431	6.56	58.60
20	442	6.73	65.33
21	388	5.91	71.23
22	377	5.74	76.97
23	345	5.25	82.22
24	276	4.20	86.42
25	223	3.39	89.82
26	167	2.54	92.36
27	159	2.42	94.78
28	118	1.80	96.58
29	82	1.25	97.82
30	59	0.90	98.72
31	36	0.55	99.27
32	26	0.40	99.67
33	12	0.18	99.85
34	5	0.08	99.92
35	4	0.06	99.98
36	1	0.02	100.00

Table B2: Frequency distribution for the grammar assessment (NPGA)

Score	Frequency	Percentage	Cumulative percentage
0	4	0.06	0.06
1	9	0.13	0.19
2	35	0.50	0.69
3	53	0.76	1.46
4	99	1.43	2.88
5	143	2.06	4.94
6	165	2.38	7.32
7	218	3.14	10.46
8	212	3.06	13.52
9	223	3.21	16.73
10	282	4.06	20.80
11	254	3.66	24.46
12	227	3.27	27.73
13	285	4.11	31.83
14	282	4.06	35.90
15	266	3.83	39.73
16	260	3.75	43.48
17	308	4.44	47.92
18	299	4.31	52.23
19	334	4.81	57.04
20	334	4.81	61.85
21	342	4.93	66.78
22	369	5.32	72.10
23	350	5.04	77.14
24	347	5.00	82.14
25	329	4.74	86.89
26	313	4.51	91.40
27	224	3.23	94.62
28	177	2.55	97.18
29	145	2.09	99.27
30	51	0.73	100.00

Further appendices

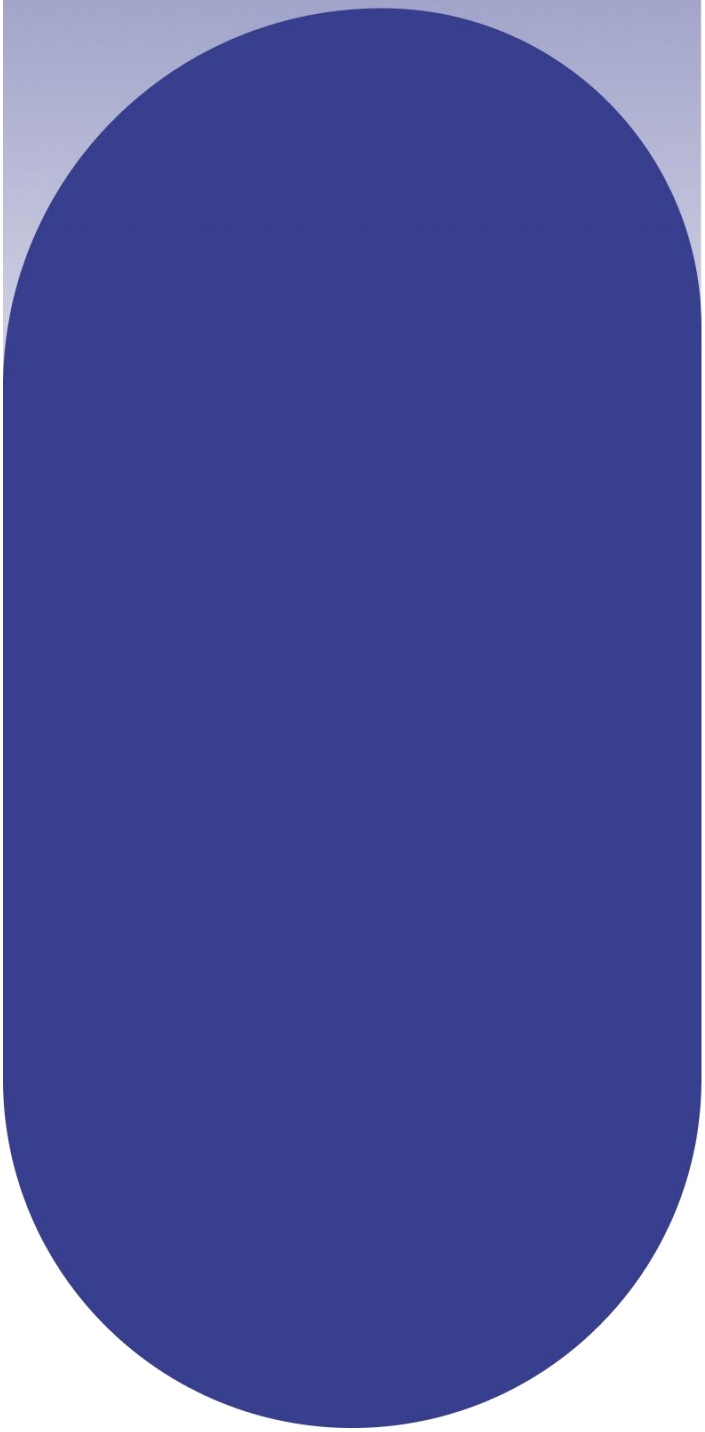
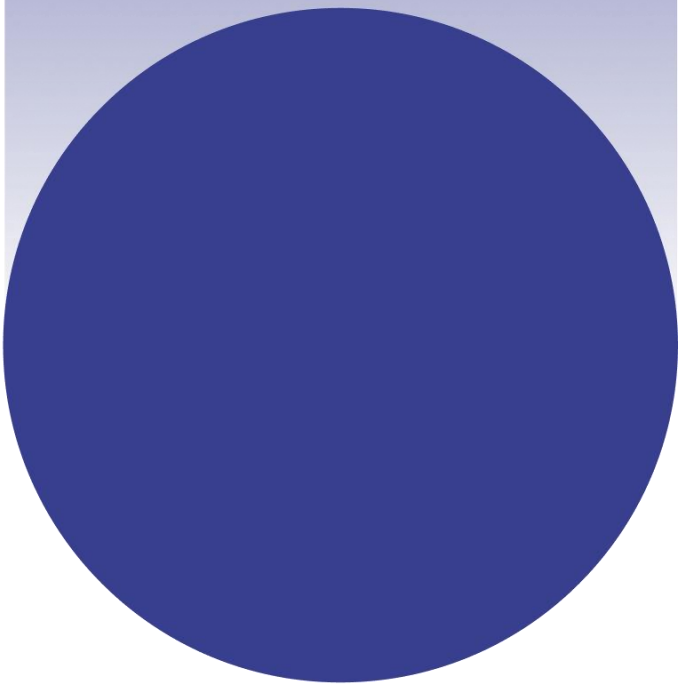
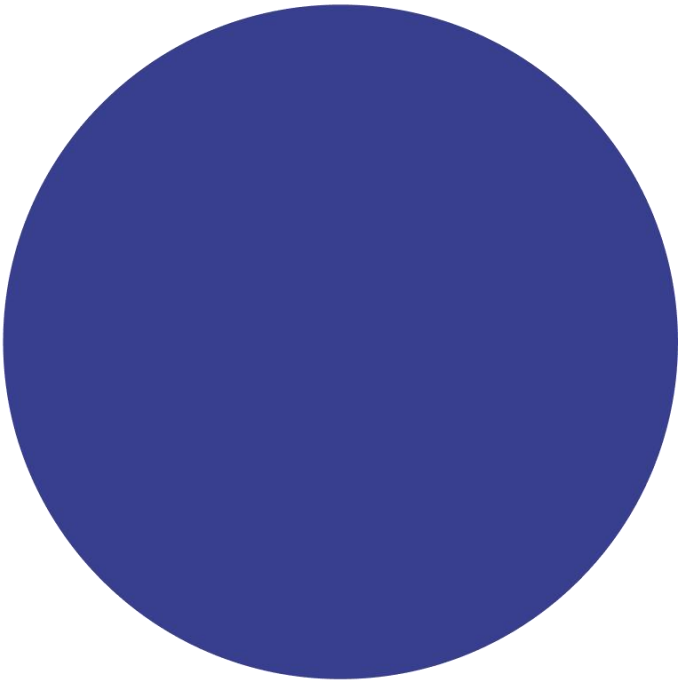
Please find the *'further appendic*es' in an accompanying document published on the EEF website.

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