

Project Title: ChatGPT in lesson preparation – A Teacher Choices Trial Evaluation Study plan

Evaluator (institution): National Foundation for Educational Research (NFER)
Principal investigator(s): Helen Poet

Evaluation summary

Project title	ChatGPT in lesson preparation – A Teacher Choices Trial
Evaluator <i>(Institution)</i>	National Foundation for Educational Research
Principal investigator(s)	Helen Poet
Study plan author(s)	Palak Roy, Katherine Aston, Ruth Staunton, David Thomas, Dr Stephen Welbourne and Helen Poet
Trial design	Two-arm cluster randomised controlled trial (RCT) with allocation at school level and teachers as study units within clusters
Trial type	Teacher Choices RCT
Pupil age range and Key stage	Ages 11-13 (Years 7 & 8)
Number of schools <i>(at design stage)</i>	58
Number of teachers <i>(at design stage)</i>	174 Y7 and Y8 Science teachers
Number of pupils <i>(at design stage)</i>	N/A
Primary outcome measure and source	RQ1: Total hours spent in lesson and resource preparation over the second five-week period measured by a weekly teacher diary.
Baseline measure for primary outcome and source	Total hours spent in lesson and resource preparation for a typical week before randomisation as measured by a teacher baseline questionnaire.
Secondary outcome measure and source	<p>RQ2: Total hours spent in lesson and resource preparation over the first five-week period measured by a weekly teacher diary.</p> <p>RQ3: Quality of lesson and resource materials used in the second five-week period measured via teachers' lesson resources ranked by an independent panel of teachers.</p> <p>RQ4: Proportion of science lessons over a five-week period where ChatGPT was used for lesson and resource preparation measured by a weekly teacher diary.</p> <p>RQ5: Proportion of weeks in each five-week period when the ChatGPT teacher guide was consulted at least once measured by a weekly teacher diary.</p>

Study plan version history

Version	Date	Reason for revision
1.0 [original]		N/A

Contents

Evaluation summary	1
Study plan version history	2
Study rationale and background	4
Teacher Choice approaches	6
Impact evaluation design	12
Implementation and process evaluation (IPE) design	27
Ethics and registration	38
Data protection.....	38
Personnel.....	41
Study Advisory Board	42
Risks	42
Timeline	44
References.....	45
Appendix A.....	47
Appendix B.....	49
Appendix C	55
Appendix D	59

Tables

Table 1: Intervention Description	8
Table 2 Trial design	12
Table 3 Sample size calculations.....	18
Table 4: Number of schools randomly allocated to each treatment arm by school size.....	19
Table 5: IPE methods overview	31
Table 6: Key members of the study team	41
Table 7: Evaluation risk assessment.....	42
Table 8: Timeline of activities for the trial	44

Figures

Figure 1: Logic model	11
Figure 2 - Participant flow diagram	19

Study rationale and background

Generative AI (GenAI) in one of its most well-known forms (ChatGPT¹) has only been available to the public since November 2022 and already has more than 100 million weekly active (global) users (Malik, 2023). ChatGPT is a large language model which generates human-like text responses to questions or prompts entered by users. It has been trained on data from the internet including websites, books, articles and manuals which it uses to predict the next word in a sequence. It has been designed to respond in an accessible and conversational manner allowing users to engage in natural language interactions on various topics and almost anyone can interact with the program once they have a log in. ChatGPT 3.5 is free to use although users can pay for a more advanced version (ChatGPT 4 Plus). Since starting the study, the free version is now updated to offer limited access to GPT-4o features which allows advanced data analysis, file uploads, vision (being able to interpret a mixture of imagery sets), web browsing, and custom GPTs (to be able to create a tailored version of ChatGPT). Whereas the Plus version allows all the advance capabilities of GPT-4o as well as enables image generation and creating and using custom GPTs (ChatGPT, 2024) . The development in this area is fast-moving and all the references in paragraph are current at the time of writing the study plan.

The Department for Education (DfE) in England recognised that the education sector was using GenAI with increasing regularity and issued a call for evidence on the topic in 2023 (DfE, 2023). It showed that these tools are already being used for lesson planning, creating resources, and writing exam questions. Benefits of using GenAI were reported to include ‘freeing up teacher time, providing additional educational support, including for pupils and students with special educational needs and disabilities (SEND) and pupils and students for whom English is an additional language (EAL), and subject specific applications’.

Polling from Teacher Tapp in September 2023 (Whittaker, 2023) suggests that a third of teachers use AI tools in their work and science teachers are among the most keen. By November 2023, Teacher Tapp was reporting 42% of teachers were using GenAI tools to help with their school work (Fletcher-Wood, 2023). A recent report (UNESCO, 2023) highlights the potential benefits of GenAI for teachers while emphasising the continued importance of skilled professionals who can prompt, appraise, refine and use outputs from GenAI programmes. However, it also warns that “despite its fluent and impressive output, GenAI cannot be trusted to be accurate.”

Although the Education Secretary has said that “AI will have the power to transform a teacher’s day-to-day work” (Keegan, 2023) , there is limited research on how teachers are actually using AI.

In Autumn 2023 the Hg Foundation and Bain & Company’s Social Impact Practice team created a guide to using GenAI in teaching. This work involved consulting with teachers and tutors about their use of GenAI and teaching practices. This culminated in the publication of a freely available web-based ‘Teaching with ChatGPT guide’ for teachers about how to use ChatGPT in their work. The guide can be found here: <https://teachingwithchatgpt.org.uk/>. The scoping and development of this web-based teacher guide was funded and supported by the Hg Foundation – who help under-represented groups to access high quality jobs in tech by supporting education- and employment- based programmes. Supporting achievement in STEM is an important part of their mission.

¹ [ChatGPT \(openai.com\)](https://openai.com/): GPT stands for Generative pre-trained transformers

About this Teacher Choices Trial

The Teacher Choices trials usually have an initial scoping phase in which the research area of interest and potential methods are investigated by the evaluation team. This Teacher Choices trial is slightly different from other Teacher Choices trials in that some scoping work² had been carried out in Autumn 2023 by Bain & Company (see above). Therefore, due to the topical nature of the choice and because the web-based teacher guide for teachers about how to use ChatGPT had already been developed, the Education Endowment Foundation (EEF) opted to move to a full Teacher Choices trial immediately, with a focus on the guide produced by Bain & Company and Hg Foundation.

This trial is jointly funded by the Education Endowment Foundation and the Hg Foundation.

This is a two-armed randomised controlled trial. The randomisation is at school level which is stratified by school size (i.e., the number of participating teachers per school) to ensure balance across the two arms. Individual/teacher level randomisation was considered but not implemented due to the risk of contamination within schools, particularly for schools where collaborative lesson planning is prevalent. Participating science teachers will be asked to implement the choices for 10 weeks over the summer term 2024 with their Year 7 and/or Year 8 classes. The two arms are: **ChatGPT and Non-GenAI** (see below for more information). The rapidly changing practices of using GenAI in education and the desire to produce robust evidence quickly, encouraged the trial commencement and completion in the academic year 2023-24.

We anticipate that in the early stages using ChatGPT may increase preparation time, but that once teachers become familiar with it, preparation time will be reduced. To cater for this, delivery is divided into two five-week periods, before and after May half-term. The overall implementation timeline of 10 weeks matches the spirit of nimble Teacher Choices RCTs, and the requirement from EEF that implementation completes within the 2023-24 academic year. The division of two blocks of five weeks is based on evidence from Bain's pilot that first five weeks block is sufficient time for teachers to look at, and try out, different parts of the teacher guide, and use ChatGPT in the second five-week block so that meaningful detectable changes in a proximal outcome can be observed.

We considered various methods for measuring impact, including on pupil outcomes. The initial request from EEF for proposals asked for the primary outcome to be pupil attainment. Our initial power calculations explored this possibility using a standardised science assessment. On this basis we calculated that we would need a sample of 156 schools to achieve an MDES of 0.2. Given only around 33% of the test material would relate to the topics taught during the delivery period (summer term), this effect would be further diluted and as a result, we would need to aim to detect an MDES of 0.07. In addition to this, given the short implementation timeline, there was a more realistic prospect of an informative proximal outcome which is more suitable to Teacher Choices trials (i.e., teacher outcomes in this instance), hence the focus of the impact evaluation is on detecting meaningful shifts in teacher preparation time and measuring lesson quality while the implementation and process evaluation (IPE) will explore perceived effects on pupil outcomes. (All information about the calculations in relation to the outcome of teacher workload can be found in the Sample size section, below.)

The main research questions are:

² Note that the published output of the scoping work was the web-based teacher guide. There was no separate scoping report on findings though Bain & Company and the Hg Foundation contributed to the set up phase for this evaluation.

Impact

RQ1 (Primary RQ): What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons over five weeks, after five weeks of initial use?

RQ2: What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons during the initial five week learning period?

RQ3: What is the effect of using ChatGPT on the quality of lesson and resource materials used in Year 7 & 8 science lessons?

RQ4: When encouraged to use ChatGPT for lesson and learning resource generation, what proportion of lessons do teachers use ChatGPT to help with preparation, and does this proportion change significantly as teachers become more familiar with ChatGPT?

- RQ5: When supplied with the teacher guide on using ChatGPT for lesson and learning resource generation, in how many weeks do teachers consult the teacher guide at least once a) during the first five weeks? b) during the second five weeks?

Implementation and process evaluation

RQ6: To what extent do teachers adhere to their allocated approaches?

RQ7: How is ChatGPT used by science teachers while preparing for lessons?

RQ8: How do teachers use the teacher guide?

RQ9: What is the perceived impact of using ChatGPT in lesson and resource preparation?

RQ10: To what extent do moderators affect behaviour and workload changes?

RQ11: What is usual practice in science teachers' lesson and resource preparation?

Appendix A summarises these research questions, including approaches to data collection and analyses.

Teacher Choice approaches

We define lesson and resource preparation as the short-term planning done by teachers to select, develop and adapt existing curriculum planning and materials in order to deliver lessons. We expect the scope and use of curriculum planning and materials to vary significantly across schools, with a minimum expectation that teachers draw on a departmental scheme of work which sets out learning objectives for each lesson.

The trial is focussed on ChatGPT as this is the GenAI tool that the teacher guide (developed by Bain & Company's Social Impact Practice team and Hg Foundation, see above) was specifically developed for. It is also worth noting another factor in our decision was that we would not be able to control for the differences between different tools when exploring GenAI vs nonGenAI (tools which function in similar ways but are based on different algorithms) and so we focus on one specific tool for the GenAI group (i.e. ChatGPT).

We carefully considered what the other group should be and decided that the most useful comparison for teachers at this early stage in the evidence base would be an arm where no GenAI tools are used at all because this condition could be considered the equivalent of business as usual prior to the public release of ChatGPT and other GenAI tools. We note that a non GenAI group is not, however, 'business as usual' in 2024, because many teachers are already experimenting with and using GenAI tools in their work. The baseline survey will describe science teachers' lesson and resource preparation before the trial, including planning workload, the extent of centralised/collective resources, preparation activities and sources. It will also describe teachers' previous use of GenAI, including for lesson planning. We will monitor any use of GenAI in the Non-GenAI group through the weekly diaries (please see the IPE section for further details).

We recognise that future research may focus on comparing different types of Gen AI and/or broadening the GenAI group to include other GenAI tools.

Brief trial guidance including Do's and Don't's for the two arms was produced by the research team and shared with teachers at randomisation (Appendix B).

ChatGPT arm

For the ChatGPT arm, the trial guidance included a link to the 'teaching with ChatGPT' website³. This website was developed by Bain & Company's Social Impact Practice in Autumn 2023 as part of a separate piece of work funded by the Hg Foundation (also see above). During the development of the website Bain worked with teachers and tutors from different school types and different subjects to explore and test out how teachers can and should use ChatGPT. This work resulted in six main 'use cases' that the guide is structured around:

- Find activity ideas
- Get ready-made practice questions
- Adapt your materials to work for your group
- Craft model answers & build mock exam questions
- Get effective explanations & step-by-step examples
- Test student understanding & avoid misconceptions

The website also explains how to start using ChatGPT, how to set up custom instructions and gives an overview of other ways teachers can use ChatGPT. It carries a warning to users to check the materials produced by ChatGPT as it can make mistakes.

Although the website does mention that personal (e.g. pupil) data should not be entered into ChatGPT, the evaluation team also reiterated this at the top of the trial guidance supplied to the ChatGPT arm due to the importance of data protection. This website is referred to as the 'teacher guide' throughout this document.

Teachers in the ChatGPT arm are asked to use ChatGPT to support their lesson and resource preparation for 10 weeks over the summer term 2024. They are asked to refer to the 'teacher guide' (the website) prior to lesson and resource preparation for these lessons. The trial guidance explains that they should use the first five weeks of the trial to look at, and try out, different parts of the teacher guide. They are encouraged to use ChatGPT in their resource preparation as much as they can during the first five weeks. After this, they can continue to refer to the teacher guide as much or as little as they would like during the second block of five weeks in the summer term. Teachers are asked not to use any other GenAI tools for their lesson and resource preparation as part of the trial relates to the web-based teacher guide, which is specific to ChatGPT.

³ <https://teachingwithchatgpt.org.uk/>

Non-GenAI arm

For the Non-GenAI arm, the trial guidance asks them specifically not to use any GenAI tools (such as ChatGPT, Gemini, or teaching specific AI tools) for their lesson and resource preparation for 10 weeks in the summer term 2024. They are asked to follow these guidelines when preparing for their Year 7 and/or Year 8 science lessons. They are allowed to draw on other sources they already use e.g., other teachers, departmental shared resources, textbooks, external schemes of work, teacher websites/forums.

The TIDieR framework for this trial is outlined in Table 1.

Table 1: Intervention Description

NAME	ChatGPT	Non-GenAI
WHY (RATIONALE)	ChatGPT is one of the most commonly recognised and used GenAI tools. It is free to access and while there are reports of it being used by teachers, little is understood about how it is being used and the impact of its use. As the teacher guide (website) had been produced specifically about how to use ChatGPT (rather than other GenAI tools), this project is focussing on ChatGPT.	Although not a 'business as usual' in 2024, this may be considered business as usual from before GenAI was made freely available. This group is asked to abstain from using any GenAI tools to minimise contamination.
WHO (RECIPIENTS)	Teachers of Year 7 and/or Year 8 science classes in English schools – the outcome of interest is their workload in preparing for lesson resources (rather than pupil attainment as it usually the case in EEF trials).	
WHAT (MATERIALS)	Brief trial guidance supplied to participating teachers, that includes Do's and Don'ts for this arm and a link to guide for teachers on how to use ChatGPT (https://teachingwithchatgpt.org.uk/) Access to ChatGPT 3.5 (free to access ⁴)	Brief trial guidance supplied to participating teachers including Do's and Don'ts for this arm.
WHAT (PROCEDURES)	Teachers are asked to use ChatGPT when preparing for lessons and creating resources (we have assumed this is based on a scheme of work or lesson plan that is already used in the school) The teacher guide emphasises that teachers do not need to create any additional lesson resources or do any additional planning specifically for this project, over and above what they usually would. All teachers are asked to teach their lessons as normal.	Teachers are asked to not use any GenAI tools (ChatGPT or otherwise) when preparing for lessons and creating resources (we have assumed this is based on a scheme of work or lesson plan that is already used in the school). The teacher guide emphasises that teachers do not need to create any additional lesson resources or do any additional planning specifically for this project, over and above what they usually would. All teachers are asked to teach their lessons as normal.

⁴ When choosing which version of ChatGPT for the trial, we aimed to replicate a 'Teacher Choice' that is easily accessible and free of cost, ensuring it would not be a financial burden. At the time of designing the trial, ChatGPT 4.0 was known to offer more advanced functionality and more sophisticated responses. We wanted to avoid a situation where some participants were testing a more advanced version than others in a non-random manner. Therefore, the free version was selected for the trial.

NAME	ChatGPT	Non-GenAI
HOW (DELIVERY GUIDE)	After randomisation, teachers will receive the trial guidance for their randomly allocated approach so that they can familiarise themselves with that approach.	
WHERE (LOCATION)	The trial will be delivered in secondary schools in England.	
WHEN & HOW MUCH (DOSAGE)	The trial will take place in the summer term 2024 (April – July 2024). Teachers are asked to use ChatGPT whilst preparing for science lessons that they deliver for the ten-week trial period and are encouraged to use as much as they can during this time.	The trial will take place in the summer term 2024 (April – July 2024). Teachers are asked not to use any form of GenAI to prepare for science lessons that they deliver during the ten-week trial period.
TAILORING (ADAPTATION)	Teachers are asked not to amend the amount or type of preparation they do for the lessons artificially for this trial. Teachers can follow the teacher guide on the (https://teachingwithchatgpt.org.uk/) website. The website recommends use of ChatGPT for six ‘use cases’ as and when relevant. These are: find activity ideas, get ready-made practice questions, adapt materials, craft model answers & build mock exam questions, get effective explanations and step-by-step examples and test student understanding and avoid misconceptions. In addition to this, teachers may find other ways to use ChatGPT in their preparation.	Teachers are asked not to amend the amount or type of preparation they do for the lessons artificially for this trial. In practice many teachers may continue to prepare for lessons in the same way they already did and using the same sources of information. However, others may use new sources of information during the trial, which, so long as they are not GenAI tools, is acceptable.

We constructed a theory of change for the ChatGPT arm of the trial, illustrated in the logic model shown in Figure 1. We used the COM-B approach (Michie, van Stralen and West, 2011) to identify the inputs (COM), that lead to the behaviour change (*B: the use of GenAI, in this case ChatGPT*). The inputs are characterised as follows:

- teachers have guidance and/or support to use GenAI (ChatGPT) when preparing lessons (capability⁵)
- teachers have access to GenAI (ChatGPT) e.g. including hardware and software, no firewall or other access restrictions at school (opportunity)
- teachers are willing to use GenAI (ChatGPT) when preparing to teach lessons (motivation).

We recognise that as part of this trial, because we are recruiting willing schools from across England, that the schools and teachers taking part in this trial are more likely to be interested and positively disposed to using GenAI/ChatGPT. This means that, for example, issues around ‘opportunity’ described above are less

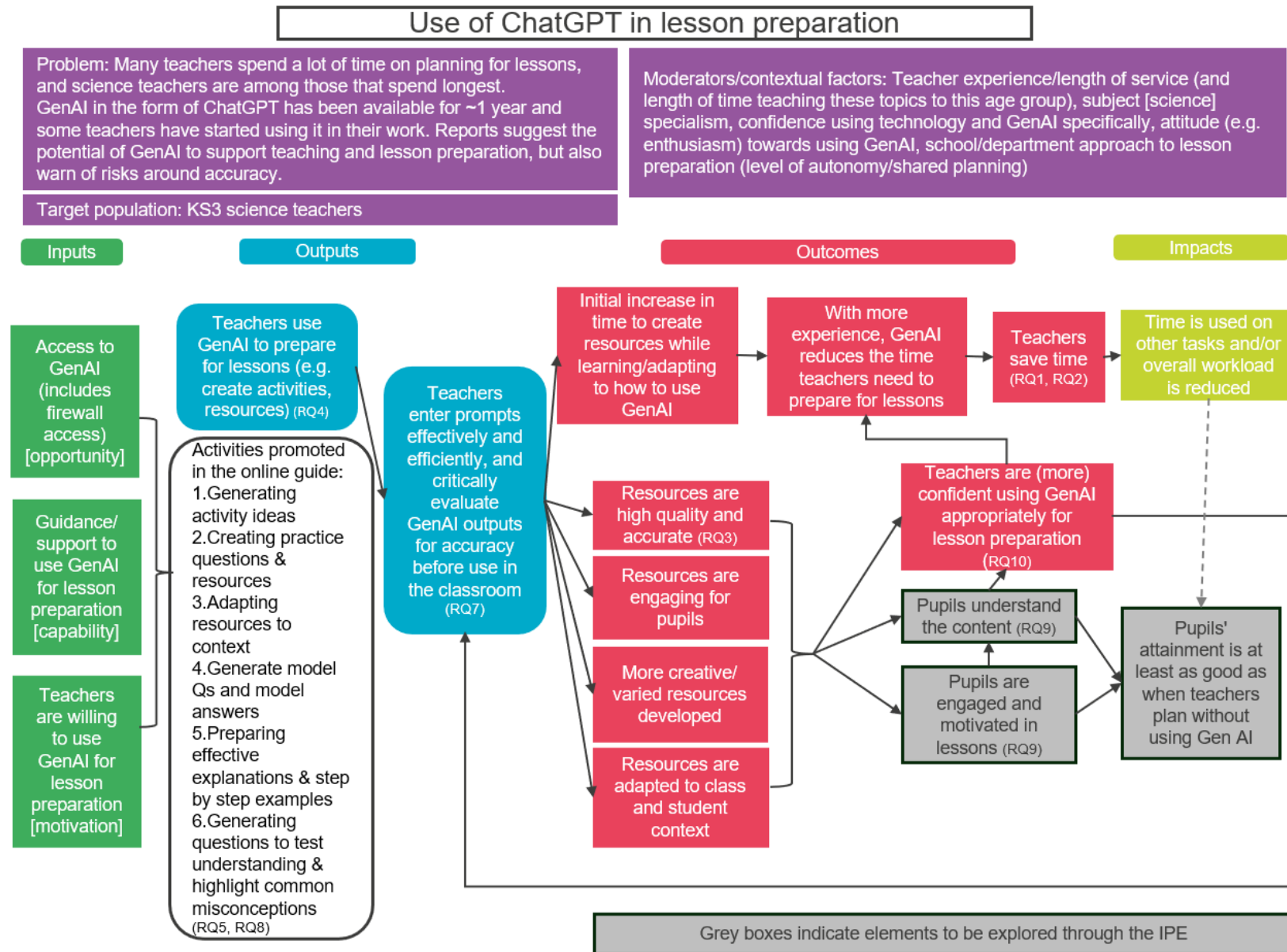
⁵ In this trial, guidance is the teacher guide available at <https://teachingwithchatgpt.org.uk/>

likely to be an issue in our sample than may be the case in the wider population, and 'motivation' to adhere to their assigned condition is likely to be high in the ChatGPT group⁶.

As illustrated in our logic model, using ChatGPT to help with lesson planning and preparation is theorised to help improve pupil outcomes via two distinct routes: by improving the quality and variety of lessons that are planned, or by freeing time which can be utilised elsewhere to improve pupil outcomes (explored in IPE RQ9). We considered using pupil outcomes as the main measure but if we selected a standardised test as the outcome, only a small part of the test would cover topics taught during the study period. Therefore, it would not give us sufficient power to detect the effect on pupil progress. Sample size calculations suggested that it would be more feasible to detect meaningful shifts in teacher preparation time, but it was very unlikely that we would be able to detect any changes in pupil attainment achieved over so short a period. A longer delivery period would be needed to test whether reduction in teacher preparation time and/or improved learning resources have the potential to result in at least as good as or improved pupil outcomes. We therefore concluded that the most appropriate and efficient method for this trial was to measure lesson quality and preparation time directly via analysis of lesson planning and teacher workload diaries. The implementation and process evaluation will examine the perceived difference in pupil engagement and response between lessons using ChatGPT and those prepared without using any GenAI.

⁶ Relatedly, due to this there may be demoralisation effects in the non-GenAI group as they are asked to refrain from using ChatGPT and other GenAI.

Figure 1: Logic model



Impact evaluation design

Research questions

- RQ1 (Primary): What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons over five weeks, after five weeks of initial use?
- RQ2: What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons during the initial five week learning period?
- RQ3: What is the effect of using ChatGPT on the quality of lesson and resource materials used in Year 7 & 8 science lessons?
- RQ4: When encouraged to use ChatGPT for lesson and learning resource generation, what proportion of lessons do teachers use ChatGPT to help with preparation, and does this proportion change significantly as teachers become more familiar with ChatGPT?
- RQ5: When supplied with the teacher guide on using ChatGPT for lesson and learning resource generation, in how many weeks do teachers consult the guide at least once a) during the first five weeks? b) during the second five weeks?

Design

Table 2 Trial design

Trial design, including number of arms		Two-arm cluster randomised controlled trial
Unit of randomisation		School
Stratification variables		School size i.e., number of participating teachers per school. Two categories: (i) 1 to 6 teachers and (ii) 7 or more teachers
Primary outcome	Variable	RQ1: Total hours spent in lesson and resource preparation over the second five-week period
	Measure (instrument, scale, source)	Researcher-designed weekly teacher diary
Secondary outcome(s)	Variable(s)	RQ2: Total hours spent in lesson and resource preparation over the first five-week period RQ3: Quality of lesson and resource materials used in the second five-week period RQ4: Proportion of science lessons over a five-week period where ChatGPT was used for lesson and resource preparation RQ5: Proportion of weeks in each five-week period when the ChatGPT teacher guide was consulted at least once
	Measure(s)	Researcher-designed weekly teacher diary (for RQ2, RQ4 and RQ5)

	(instrument, scale, source)	Rank order for the lesson and resource material, rated by the teacher panel (RQ3)
Baseline for primary outcome	Variable	Total hours spent in lesson and resource preparation for a typical week before randomisation
	Measure (instrument, scale, source)	Researcher-designed teacher survey prior to randomisation
Baseline for secondary outcome	Variable	Baseline for RQ2: Total hours spent in lesson and resource preparation for a typical week before randomisation Baseline for RQ4: Use of GenAI before randomisation RQ3 and RQ5 analyses will not have baseline measures.
	Measure (instrument, scale, source)	Researcher-designed teacher survey prior to randomisation

This study is a cluster-randomised controlled Teacher Choices trial with the randomisation at school level. Schools have been randomly allocated to one of two groups: ChatGPT group and Non-GenAI group. The implementation lasts for ten weeks during the summer term of 2024. Teachers were asked to nominate their Year 7 and/or Year 8 classes for this trial prior to randomisation. The two arms are as follows:

ChatGPT group: science teachers in this group are asked to use ChatGPT to prepare lessons and resources for upcoming Year 7 and/or 8 science lessons. They also receive access to the online ChatGPT guide to guide their lesson and resource preparation.

Non-GenAI group: science teachers in this group are asked not to use ChatGPT or any other GenAI tool in any lesson and resource preparation for their Year 7 and/or 8 science lessons.

Once the school signed up to the trial, teachers were asked to respond to a baseline survey. Questions in this survey form the baseline measures for several outcomes hence the survey completion was a requirement to be part of the trial. After the baseline surveys were completed (surveys closed on 19th March 2024), schools were randomly allocated to one of the two groups. Teachers were notified about their group allocation on 25th March 2024. Along with this, we also sent a document which gave an overview of their allocated approach to include brief Do's and Don'ts for the approach (see Appendix B). Teachers were asked to use their allocated approach for ten weeks in the summer term 2024.

Data collection activities for both randomised groups during implementation include:

- In each of the ten weeks of delivery, teachers will be asked to complete a short online diary entry regarding the lessons they delivered in that week. This will form the primary and secondary outcomes. (ChatGPT and Non-GenAI groups)
- After the first block of five weeks, just before the May half term, teachers in the ChatGPT arm will be asked to complete a short, multiple-choice quiz. (ChatGPT group only)
- We will also conduct case study visits to 12 schools which will include a senior leader interview, teacher focus group, lesson preparation walkthrough and pupil focus groups. (ChatGPT and Non-GenAI groups)

- Teachers in the ChatGPT group will be asked to submit their ChatGPT transcript⁷. (ChatGPT group only)
- Teachers in both groups will also be asked to submit lesson resources for three lessons that they delivery during the second block of five weeks. (ChatGPT and Non-GenAI groups)
- At the end of the ten-week period, teachers will also be asked to complete a short endpoint teacher survey. (ChatGPT and Non-GenAI groups)

Schools that complete all ten of the weekly diaries will receive a payment of £100 per participating teacher at the end of the ten-week trial. In addition to this, schools will also receive £30 per teacher where teachers submit lesson resources for three lessons.

The trial focuses on measuring how much time teachers spend preparing for lessons and assessing the quality of their lesson materials over any other pupil outcomes as the pupil outcomes are considered longer term and may take a while before an effect is detected. The trial's **primary outcome is teacher workload**, specifically the time spent on lesson and resource preparation for Year 7 and/or 8 science lessons over five weeks, after five weeks of initial use. There are a number of **secondary outcomes**:

- total hours spent on lesson and resource preparation over the first five-week period,
- quality of lesson and resource materials used in lessons delivered in the second five-week period,
- proportion of science lessons over a five-week period where ChatGPT was used for lesson and resource preparation,
- proportion of weeks in each time period (weeks 1–5 and weeks 6–10) when the ChatGPT guide was consulted at least once by teachers in the ChatGPT group.

While the impact evaluation looks at teachers' workload in lesson and resource planning as well as the quality of the materials, IPE will explore what the saved time is used towards. For example, whether it represents a reduction in overall workload or whether it is diverted elsewhere. There may also be some differences depending on teacher characteristics such as whether they are early career or more experienced teachers or their prior use of GenAI etc. (see IPE RQ9 for more information).

See the outcomes section for further details.

In addition, as mentioned above, all teachers in the ChatGPT group will be asked to complete a short quiz (8 multiple-choice items, possible score range 0-8) which assesses their understanding of the ChatGPT teacher guide (website) content. This quiz aims to encourage teachers to engage with the guide and to check teachers' awareness of the principles and recommended practice set out in the guide. The quiz was developed by Bain & Company and NFER. The quiz will be undertaken towards the end of the five-week 'learning period' (May 2024), to assess teachers' awareness after opportunities to engage with the teacher guide and use ChatGPT when preparing lessons, and to check awareness of key principles at the end of

⁷ ChatGPT transcript is the text exchange between the user and the ChatGPT, comprising the instructions/requests they have sent to ChatGPT, and the response/output generated by ChatGPT. ChatGPT functionality allows the user to export and share this. Please see the IPE section for further details.

the five-week 'learning period'. The quiz will be used for the compliance measure (see the analysis section for further details).

Participant selection

School eligibility

Any state secondary school in England can take part as long as there is at least one teacher who teaches Years 7 and/or Year 8 science, and who completes the baseline survey.

Teacher eligibility

Any science teacher who teaches Years 7 and/or Year 8 science (including non-specialist science teachers and Early Career Teachers (ECT) who is willing to be part of the trial. Teachers are only considered eligible once they complete the baseline teacher survey for the trial.

Recruitment

NFER was responsible for school recruitment for this trial. The power calculations at design stage suggested that the trial requires 174 teachers and 58 secondary schools (see the sample size section for further details). Recruitment activities took place between 1st February and 12th March 2024. EEF and NFER promoted the trial on their respective websites and social media platforms. Interested teachers and leaders expressed initial interest in the trial by filling out an online Expression of Interest (EOI) form. NFER contacted the individuals who signed the EOI and asked them to ask the headteacher or the senior leader at their school to complete the Memorandum of Understanding (MoU). Soon after the MoU was signed, schools were asked to submit information about participating teachers. This included names, contact details and the number of Year 7 and/or 8 science classes they taught. Once this was received, NFER sent baseline surveys to all these teachers. The baseline survey completion was a requirement for individual teacher participation and hence a criterion for a school to be part of the trial. The baseline surveys were completed between 28th February and 19th March 2024.

Outcome measures

Baseline measures

The baseline measure for RQ1 and RQ2 will be teachers' workload in lesson and resource preparation for a typical week prior to randomisation. This is measured via a workload question in the baseline teacher survey where teachers indicated the number of hours they spent on lesson or resource preparation for their Year 7 and/or Year 8 science lessons during their most recent complete calendar week. This measure includes all activities conducted by teachers towards lesson and resource preparation outside of regular class hours, including weekends and evenings, while excluding their time spent on teaching, marking, or administrative tasks. This survey question is based on questions used in previous workload studies (OECD, 2018, Walker, Worth and Van den Brande, 2019) although due to our particular focus of lesson and resource preparation we only asked about that domain, whereas previous studies asked about a wider range of workload domains (e.g. marking and administration). While the baseline measure is workload in a single week and the outcome measures for RQ1 and RQ2 are cumulative workload over five weeks, we do not anticipate that the difference in scale will meaningfully reduce the baseline measure's potential to

explain variability in the outcome measures, since both are still measures of workload, the baseline is close in time just before randomisation, and each teacher's individual baseline and outcome data will be linked in the analysis models.

The baseline for RQ4 is a measure of GenAI use prior to randomisation. This is measured by a question in the teacher baseline survey which asks whether the respondents have used ChatGPT or any other GenAI tool for any purpose. This measure will be an ordinal variable to indicate the extent of GenAI familiarity and use at baseline, with categories 'Yes, frequently', 'Yes, occasionally', 'Yes, but only once or twice', 'No, never', and 'Not sure'.

There are no baseline measures for RQ3 and RQ5.

Primary outcome

The primary outcome for this trial will be teachers' workload in lesson and resource preparation over the second five-week period of the trial. This is measured via a workload question in the weekly teacher diary. It will be a single data point per teacher calculated as the cumulative hours dedicated to lesson and resource preparation for their nominated Year 7 and/or Year 8 science lessons over the second five-week period (weeks 6–10 of the trial) during the second half of the summer term 2024 i.e., the sum of five values for each diary respondent. Up to two missing values per five-week block will be imputed as the mean of the other values for that teacher. This measure will include all activities conducted by teachers towards lesson and resource preparation outside of regular class hours, including weekends and evenings, while excluding their time spent on teaching, marking, or administrative tasks. The diary question used to calculate the primary outcome measure is based on previous workload studies and the question used at baseline although there is a slight difference. In the diary, the question is about the time spent preparing for lessons that took place in a specific week (rather than preparation that took place in that week as asked in the baseline survey) (see Q2 in Appendix C). This question is repeated every week of the ten-week trial period.

Prior to analysis, we will check the for the presence of 'extreme' values given in this question. Responses to numeric questions in the TALIS workload survey (OECD, 2018) are excluded from analysis if they are deemed implausible. For the workload questions on which we modelled our primary outcome, implausible is defined as values larger than 120 hours, which are excluded. Collection methodology in our study already ensures that workload values input for each week cannot be below 0 or above 99 hours. The frequency of values outside of the range defined by the mean \pm 3.29*standard deviation (Tabachnick and Fidell, 2013) will be reported for both randomised groups. We will retain the data for analysis as we have no evidence that they are not true data points.

Secondary outcomes

The **secondary outcome for RQ2** will be the teacher workload in lesson and resource preparation over the first five-week period of the trial. This is measured via the same workload question in the weekly teacher diary as the primary outcome measure (see above). It will be the cumulative hours dedicated to lesson and resource preparation for their nominated Year 7 and/or Year 8 science lessons over the first five-week period (weeks 1 to 5 of the trial). As for RQ1, up to two missing values per five-week block will be imputed as the mean of the other values for that teacher.

The **secondary outcome for RQ3** will be the quality of lesson and resource materials used in Year 7 and/or Year 8 science lessons delivered in the second five-week period (weeks 6–10) of the trial. The quality will be measured via a rank awarded to lesson resources supplied by teachers (n=40). The rank will

be determined by a panel of five experienced science teachers or leaders (who will be blind to group allocation). The panel will rank the lessons against each other, which means the measure will be a numerical variable with a range 1–40. For further details on the sample and ranking process, see the analysis section.

The **secondary outcome for RQ4** will be the total number of lessons where ChatGPT was used and the total number of lessons where ChatGPT was not used, in the nominated Year 7 and/or Year 8 science lessons over each five-week period. This will be measured using two questions from the weekly teacher diary – the number of lessons where ChatGPT was used and the total number of science lessons in a week (see Q3a and Q1 in Appendix C). While the outcome is a binomial success/failure variable, it will be reported as the proportion of nominated Year 7 and/or Year 8 science lessons over a five-week period where ChatGPT group teachers used ChatGPT during lesson and resource preparation.

$$Prop_{ChatGPT} = \frac{\text{no. of lessons where ChatGPT was used for resource preparation}}{\text{no. of total lessons}}$$

This proportion will be estimated from the model for two time points - weeks 1–5 and weeks 6–10 since the proportion will use the numerator and denominator that will be the total of respective values for a five-week period.

The **secondary outcome for RQ5** will be the proportion of weeks in each time period (weeks 1–5 and weeks 6–10) the ChatGPT guide was consulted at least once. We will measure this by asking a question (to ChatGPT group teachers only) in the weekly diary about whether they used the teacher guide from teachingwithchatgpt.org (see Q4 in Appendix C). Then, we will calculate the proportion of weeks they confirmed the use ("yes") out of the total weeks in each period. This means, the measure will be a tally of "yes" responses for each week, divided by the total number of weeks (five for each period). The resultant variable will be a numeric with a range 0–1 for each time period.

Sample size

The outcomes for this study are all teacher rather than pupil-related so there is no separate calculation for FSM pupils. Sample sizes were calculated using data from the TALIS workload survey (OECD, 2018) to generate the expected distribution of lesson preparation times. The TALIS workload survey contains data on secondary science teachers' weekly lesson preparation workload. We assumed that this weekly workload for all science classes (assuming a teacher teaches across five year groups, Years 7-11) would be similar to the secondary science teachers' total preparation for one year group across five weeks. The TALIS study showed that teachers who felt they spent too much time on non-teaching activities spent 2.3 hours longer (per week) planning for lessons than those who felt their time for these activities was 'about right'. Based on this we aimed to power the study to detect a two-hour change in workload for one year group across a five-week period. In the initial proposal we assumed we would be looking at only one year group (Year 8), but in the setup stages we realised we could also include Year 7 teaching to increase the power for no additional cost. Therefore, the calculations below reflect each participating teacher teaching both year groups.

Table 3 Sample size calculations. Values in brackets indicate MDES and numbers after accounting for assumed attrition.

		Design	Randomisation
Minimum Detectable Effect Size (MDES)		0.323 (0.342)	0.268 (0.297)
Pre-test/ post-test correlations	level 1 (teacher)	0.7 ⁸	0.7
	level 2 (school)	0	0
Intracluster correlation (ICC)	level 2 (school)	0.02 ⁹	0.02
Alpha		0.05	0.05
Power		80%	80%
One-sided or two-sided?		Two-sided	Two-sided
Average cluster size		3 (3)	3.8 (3.4)
Number of schools	ChatGPT group	29	34
	Non-GenAI group	29	34
	Total	58 (52)	68 (61)
Number of teachers	ChatGPT group	87	129
	Non-GenAI group	87	129
	Total	174 (156)	258 (207)

All sample size calculations were conducted using the software package PowerUpR (Bulus *et al.*, 2021).

During design-stage calculations, we chose a conservative number of teachers per school, assumed 10% school attrition and 0% teacher attrition from the number of schools and teachers shown in the table above. Randomisation stage sample size calculations assumed 10% school attrition and 10% teacher attrition from the number of schools and teachers shown in the table above. The teacher-level attrition was added after randomisation as we now know the number of teachers per school, and we do expect some teachers will not complete weekly diaries. Attrition corrected MDES and numbers of teachers/schools for both stages are shown in brackets in the table.

As can be seen from the table, we are benefitting of a lower MDES at randomisation due to increased number of schools and teachers that we were able to recruit to the trial. Using the standard deviation derived from the TALIS data, detection of a 2.3 hours difference in workload translates to an MDES of 0.299 which is very close to the MDES at randomisation after attrition is accounted for.

⁸ Due to lack of references for pre-post correlation in workload outcomes, this value is an estimate. The value found in this trial will be prominently reported to inform sample size calculations for future work in this area.

⁹ ICC estimated from the secondary science teachers' weekly lesson preparation responses in the TALIS workload survey

Randomisation

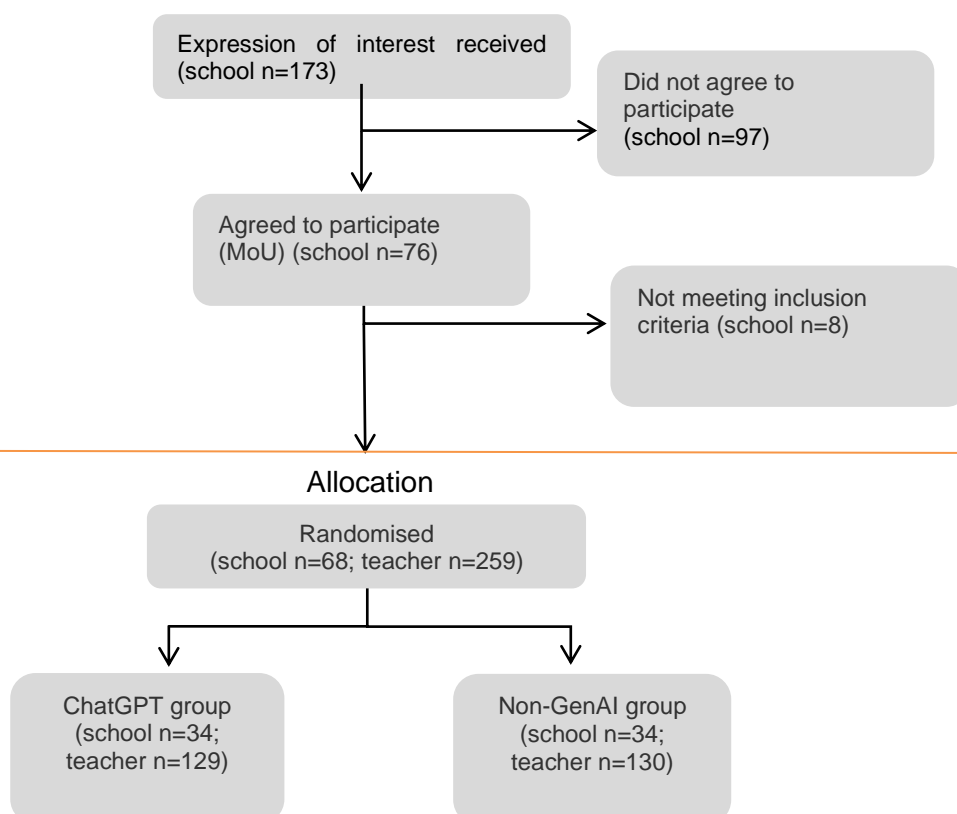
This is a cluster trial with randomisation at school-level. Schools were randomly assigned to one of two arms with equal allocations (1:1). Recruitment resulted in variation in the number of participating teachers per school, with a minority of schools (n=8, 3% of recruited schools) with a high number of teachers (between 7 and 10). We therefore stratified the randomisation by size of the school (in terms of participating teachers) in order to have similar number of teachers per arm. School size was determined by the number of teachers who complete the baseline survey. There were two strata – schools with 1-6 teachers and those with 7 or more teachers. Table 4 presents school numbers in each stratum and randomised groups.

Table 4: Number of schools randomly allocated to each treatment arm by school size

	ChatGPT	Non-GenAI
Small schools (1-6 participating teachers)	30	30
Large schools (7-10 participating teachers)	4	4

The randomisation was undertaken by an NFER statistician on 19th March 2024. Randomisation was carried out in R, and syntax scripts were saved in order to ensure transparency and replicability (see Appendix D). The randomisation process was quality assured by another statistician from NFER's Centre for Statistics. Once randomisation was completed, schools were notified of their group allocation on 25th March 2024. The participant flow until group allocation is presented in Figure 2.

Figure 2 - Participant flow diagram



Statistical analysis

Primary analysis (RQ1 - What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons over five weeks, after five weeks of initial use?)

We will use a linear multilevel model to estimate the effect of using ChatGPT on teacher preparation time in the second five-week time period, controlling for baseline lesson and resource preparation time by means of a covariate and accounting for clustering of teachers at school level. The teacher-level regression model is defined by:

$$Workload_{ij} = \beta_{0j} + \beta_1 ChatGPT_j + \beta_2 BaseWorkload_{ij} + \beta_3 StratificationGroup_j + \varepsilon_{ij}$$

Where:

- ***Workload_{ij}*** is the preparation workload of teacher i in school j for Year 7 and/or Year 8 summed across the second five-week period. This will be measured using a weekly teacher diary.
- ***BaseWorkload_{ij}*** is the estimated weekly preparation workload for years 7 and/or 8 science lessons in their most recent complete calendar week prior to the baseline survey for teacher i in school j as measured by the baseline teacher survey.
- ***StratificationGroup_j*** is the stratification group for school j. Stratification group will be determined by the number of teachers in the school completing the baseline survey (1-6, or greater than 7).
- ***ChatGPT_j*** is a variable indicating whether school j was randomly assigned to use ChatGPT or not.
- **β_{0j}** is the intercept in school j (modelled as a random effect).
- **β_1** is the coefficient of interest estimating how much difference being assigned to the ChatGPT group makes to teacher workload.
- **β_2** is a coefficient estimating the association between baseline teacher workload and workload in the second five-week period.
- **β_3** is a coefficient estimating the association between the stratification group and the workload in the second five-week period.
- **ε_{ij}** is the residual error term for teacher i in school j.

To avoid the scenario where teachers who only teach one or two lessons (or a large number of lessons) unduly influence the results we will weight the model by the total number of lessons taught to qualifying year-groups in the time period, scaled so that weights sum to one. For the primary analysis, unweighted model coefficients will also be reported as a sensitivity check, in order to demonstrate the effect that skewed delivery volume has on the results.

Distributional assumptions of the model will be assessed through visual inspection of the residual plots. Should a strong heterogeneity of variance be apparent, a transformation (\log_{10}) of the outcome variable will be applied. Model assumptions will be rechecked and if heterogeneity of variance is still apparent a non-parametric alternative method, such as a Friedman ANOVA, will be applied.

Secondary analysis

RQ2 (What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons during the initial five week learning period?) will be addressed using an identical model to that used for the primary analysis with the exception that the modelled workload will be summed across the first five weeks of the trial rather than the second five weeks. Distributional assumptions will be checked and the same decision process applied in RQ1 will be enacted.

To answer **RQ3** (What is the effect of using ChatGPT on the quality of lesson and resource materials used in Year 7 & 8 science lessons?) we will ask each teacher to submit all of their lesson resources and materials for three of their nominated Year 7 and/or Year 8 science lessons delivered in weeks 6 to 10. Our assumption is that for ChatGPT group teachers, this material will be generated using ChatGPT. There will be some variation in the amount and details of the material supplied by teachers depending on the science topic. For each of the three selected lessons, we will request all lesson and resource material including the lesson plan, overall objectives and tasks to support the appraisal. Note we expect that the supporting documents and/or schemes of work would have been generated by central resources at science departments and would not have been generated using ChatGPT; these are being gathered for context.

Once we receive these resources, we will then randomly select 40 of these (20 from each arm of the trial) subject to the constraint that resources from only one lesson can be selected from each school. If we are not able to collect lesson resources from more than 40 schools, we will raise this limit to 2 lesson resources per school, but add a new restriction that only one resource can be selected per teacher.

These lesson resources will then be evaluated by a panel of five experienced secondary science teachers who will be blind to group allocation. The panel members will be given guidance on how to review the resources prior to the panel workshop. The guidance will be developed by NFER's science assessment lead who will be blind to group allocation and the source of the lesson resources. The guidance will be reviewed by the project director and the trial manager to assure alignment with the objectives of RQ3. Each panel member will review resources for 16 lessons independently ahead of the panel workshop against four criteria giving a score out of five for each lesson resource: (i) clarity of lessons resources, (ii) to what extent do activities engage students with the learning and check their understanding', (iii) appropriateness for the age group and ability level of the class and (iv) quality and accuracy of scientific content. This way, each lesson resource will have two initial scores out of 20 given by two reviewers. In addition to the scores, the reviewers will also make supporting notes to explain the reasoning behind the scoring describing the features that attracted the scores (this will be presented during the panel workshop). NFER will create an average score for each lesson, which will be used to sort the lessons in five groups with each group consisting of levels with similar score levels (e.g. the highest scoring lessons in one group, the lowest scoring lessons in another).

These groupings will feed into a panel discussion and a ranking process that takes place during a panel workshop as described below. This will be attended by all panel members and facilitated by NFER. During the workshop, a panel discussion will take place where the panel will go through all the lessons within a group one after the other. The reviewers will present their notes and individual scores whilst the panel discusses the lessons to determine the final score for each lesson in the group. This will help to differentiate between lessons of similar scores. Where lessons have the same final score, a majority decision by the panel will determine their rank order in the group. The panel will further discuss the lessons with the highest and lowest scores within each group in case these lessons need to move up or down across the groups. Once this process is completed, lessons will be assigned ranks of 1–40 based on the

overall final scores and any individual ranking decisions. This means the outcome measure (the rank) will be a numerical variable with a range 1–40. After the workshop, the ranks will be matched with group allocation by the analyst and will be analysed using a Mann-Whitney U test to determine whether there is any difference in overall quality between lessons in each of the experimental conditions.

For **RQ4** (When encouraged to use ChatGPT for lesson and learning resource generation, what proportion of lessons do teachers use ChatGPT to help with preparation, and does this proportion change significantly as teachers become more familiar with ChatGPT?) we will build a binomial multilevel model with a logit link function to compare the proportion of lessons where ChatGPT was used to assist with preparation across the two time periods. This comparison will be restricted to the ChatGPT group controlling for baseline AI use by means of a covariate and accounting for clustering of teachers at school level. The regression model is defined by:

$$\text{logit}(p\text{ChatGPT}_{ij}) = \beta_{0j} + \beta_1\text{TimePeriod}_i + \beta_2\text{BaseAIuse}_{ij}$$

Where:

- **$p\text{ChatGPT}_{ij}$** is the probability that teacher i in school j uses ChatGPT to aid lesson preparation for Year 7 and/or 8 science lessons. This will be measured using two questions from the weekly teacher diary and parameterised as the total number of ‘successes’ and ‘failures’ across the five weeks for each teacher in each time period.
- **BaseAIuse_{ij}** is a measure of how much teacher i in school j used Generative AI prior to the study as measured by the teacher baseline survey.
- **β_{0j}** is the intercept in school j (modelled as a random effect).
- **β_1** is the coefficient of interest estimating how much frequency of ChatGPT use changes across the two time-periods.
- **β_2** is a coefficient estimating the association between baseline teacher AI use and the frequency of using ChatGPT.
- **TimePeriod_{ij}** is a variable indicating whether the frequency of ChatGPT use reported by teacher i in school j was in the first or second time-period.

For **RQ5** (When supplied with the teacher guide on using ChatGPT for lesson and learning resource generation, in how many weeks do teachers consult the guide at least once a) during the first five weeks? b) during the second five weeks?) we will not run any statistical models, but will instead report the proportion of weeks in each time period where the guide was consulted at least once (ChatGPT group only). There will be no imputation for this analysis; thus, it will be a complete case analysis.

Subgroup analysis

We will perform a subgroup analysis to explore the presence of heterogeneous treatment effects across the ‘approach to planning’ subgroups (individual lesson planning, central lesson planning). We will expand the primary analysis linear multilevel model to include an interaction with the subgroup variable, and test for

differences between ChatGPT and Non-GenAI groups within each level of ‘approach to planning’. The teacher-level regression model is defined by:

$$Workload_{ij} = \beta_{0j} + \beta_1 ChatGPT_j + \beta_2 ApproachToPlanning_{ij} + \beta_3 (ChatGPT_j * ApproachToPlanning_{ij}) + \beta_4 BaseWorkload_{ij} + \beta_5 StratificationGroup_j + \varepsilon_{ij}$$

Where:

- ***Workload_{ij}*** is the preparation workload of teacher i in school j for Year 7 and/or Year 8 summed across the second five-week period. This will be measured using a weekly teacher diary.
- ***BaseWorkload_{ij}*** is the estimated weekly preparation workload for years 7 and/or 8 science lessons for teacher i in school j as measured by the baseline teacher survey.
- ***StratificationGroup_j*** is the stratification group for school j. Stratification group will be determined by the number of teachers in the school completing the baseline survey (1-6, or greater than 7).
- ***ChatGPT_j*** is a variable indicating whether school j was randomly assigned to use ChatGPT or not.
- ***ApproachToPlanning_{ij}*** is a variable indicating which approach to planning teacher i in school j indicated more strongly¹⁰ in the endpoint teacher survey.
- ***β_{0j}*** is the intercept in school j (modelled as a random effect).
- ***β₁*** is the coefficient of interest estimating how much difference being assigned to the ChatGPT group makes to teacher workload.
- ***β₂*** is a coefficient estimating the association between approach to planning and teacher workload.
- ***β₃*** is a coefficient estimating effect of the interaction between ChatGPT group and approach to planning.
- ***β₄*** is a coefficient estimating the association between baseline teacher workload affects workload in the second five-week period.
- ***β₅*** is a coefficient estimating the association between the stratification group and the workload in the second five-week period.
- ***ε_{ij}*** is the residual error term for teacher i in school j.

To avoid the scenario where teachers who only teach one or two lessons (or a large number of lessons) unduly influence the results we will weight the model by the total number of lessons taught to qualifying year-groups in the time period.

¹⁰ Where a teacher gave equal weight to both approaches, they will be categorised as favouring individual lesson resource preparation

In addition, we will run separate models for the two levels of ‘approach to planning’. These will take the exact form of the primary analysis described for RQ1, but with only the subset of teachers in each approach to planning group.

Estimation of effect sizes

In line with the EEF statistical guidelines (EEF, 2022) the effect size for the primary outcome will be reported as Hedges g calculated using the following equation:

$$ES = (\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}} / sd_{\text{pooled}}$$

Where:

- $(\bar{Y}_T - \bar{Y}_C)_{\text{adjusted}}$ denotes ANCOVA difference in means between trial groups adjusting for baseline workload, in our models this is equivalent to the coefficient β_1 so this model coefficient will be used in the calculation.
- sd_{pooled} is the unconditional standard deviation of primary outcome measure pooled across the two groups, calculated as

$$s_p = \sqrt{\frac{(n_T-1)s_T^2 + (n_C-1)s_C^2}{(n_T-1) + (n_C-1)}}$$

Confidence intervals for each impact estimate will be estimated by adding/subtracting from the point estimate the standard errors of the ChatGPT coefficient multiplied by the left-tailed inverse of the Student’s t -distribution with a probability of 2.5% and the number of degrees of freedom associated with the size of the sample. The confidence around the impact estimates will be converted to effect size confidence intervals using the same formula as the effect sizes themselves.

Analysis in the presence of non-compliance

Compliance for this study will be defined separately for the ChatGPT and Non-GenAI groups.

Teachers in the Non-GenAI group will be considered compliant if they do not use ChatGPT or any GenAI tool for lesson and resource preparation for their nominated Years 7 and/or Year 8 science lessons throughout the ten-week period. This will be measured via a diary question where they are asked to indicate the use of ChatGPT or any other Non-GenAI use for the lessons delivered that week. This has to be zero in order for the Non-GenAI teachers to be compliant.

For teachers in the ChatGPT group, teachers are asked to read and follow the teacher guide as well as use ChatGPT when preparing for science lessons. Both these components will together form the compliance measure for this group. Teachers will be considered compliant if they:

- pass the quiz with at least 75% correct answers, and
- if they fall within the top 60% of ChatGPT arm teachers who have used ChatGPT to help with lesson preparation for the lessons delivered during weeks 6–10 of the trial. The 60th percentile will be calculated using proportions of lessons where teachers used ChatGPT throughout the second five-week period. Proportions will be calculated for those teachers who complete at least three weeks of diaries in the week 6 – 10 time period, with no imputation.

These criteria have been selected because:

we believe that more than 6 out of 8 correct answers in the quiz sufficiently demonstrates that teachers have read and understood the teacher guide, and the threshold is appropriate to account for teachers' awareness at the end of the five-week 'learning period'.

We will conduct an additional analysis among the compliant teachers to estimate the complier average causal effect (CACE). An instrumental variable (IV) analysis will be performed, using two-stage least squares methods to estimate the effect of compliance with the study guidance on lesson preparation workload. For the first stage the compliance indicator will be regressed on random assignment, together with covariates from the primary analysis model. For the second stage lesson preparation workload will be regressed on each teacher's predicted compliance value from the first stage, together with covariates from the primary analysis model. The coefficient for predicted compliance in this second stage is the CACE (complier average causal effect) estimate for the effect of compliance (as defined above) on lesson preparation workload. Results from both stages will be reported. The stage one logistic regression model is defined as:

$$p\text{Compliance}_{ij} = \text{logit}^{-1}(\beta_{10j} + \beta_{11}\text{ChatGPT}_j + \beta_{12}\text{BaseWorkload}_{ij} + \beta_{13}\text{StratificationGroup}_j)$$

The stage two regression model is defined as:

$$\text{Workload}_{ij} = \beta_{20j} + \beta_{21}\widehat{\text{Compliance}}_{ij} + \beta_{22}\text{BaseWorkload}_{ij} + \beta_{23}\text{StratificationGroup}_j + \varepsilon_{ij}$$

Where:

- ***Compliance_{ij}*** is binary indicator of compliance for teacher i in school j.
- ***Compliance_{ij}*** is the predicted probability of compliance teacher i in school j from the stage one model.
- ***Workload_{ij}*** is the preparation workload of teacher i in school j for Year 7 and/or Year 8 summed across the second five-week period. This will be measured using a weekly teacher diary.
- ***BaseWorkload_{ij}*** is the estimated weekly preparation workload for years 7 and/or 8 science lessons in their most recent complete calendar week prior to the baseline survey for teacher i in school j as measured by the baseline teacher survey.
- ***StratificationGroup_j*** is the stratification group for school j. Stratification group will be determined by the number of teachers in the school completing the baseline survey (1-6, or greater than 7).
- ***ChatGPT_j*** is a variable indicating whether school j was randomly assigned to use ChatGPT or not.
- **β_{10j}** and **β_{20j}** are the intercepts in school j (modelled as a random effect).
- **β_{11}** is the coefficient estimating how much difference being assigned to the ChatGPT group makes to probability of compliance.
- **β_{12}** is a coefficient estimating the association between baseline teacher workload and probability of compliance.

- β_{13} is a coefficient estimating the association between the stratification group and the probability of compliance.
- β_{21} is the coefficient estimating how much difference being compliant makes to workload in the second five-week period.
- β_{22} is the coefficient estimating the association between baseline teacher workload and workload in the second five-week period.
- β_{23} is the coefficient estimating the association between stratification group and the workload in the second five-week period.
- ε_{ij} is the residual error term for teacher i in school j

Additional analyses and robustness checks

We will run an additional analysis to see if the proportion of lessons for which ChatGPT is used mediates the effect on workload in the ChatGPT group.

How does the proportion of lessons for which ChatGPT is used mediate changes in teacher workload?

We will follow EEF's guidance on recommended approach to undertaking and reporting mediation analysis¹¹.

Moderator analyses are presented in implementation and process evaluation section.

Missing data analysis

If there is missing diary data (i.e. the outcome variables for RQ1 and RQ2), the reason for this will be established where possible and described in the final report. If the reasons for missing diary data can be considered independent of this trial e.g. due to staff absence missing data can safely be treated as missing at random. To investigate patterns of missingness, we will fit a binomial generalised linear mixed effects model with a binary indicator identifying 'any missing data' as the response against teacher characteristics collected in the baseline survey i.e. whether trial lead, current role, main teaching subject, number of years of teaching experience, whether trainee/ECT and ITT subject. School will be the random effect.

Where there are missing entries, we will replace up to two numeric diary questions in each time period with the mean for the teacher in the time period¹². Imputation will be required for missing diary data as the outcome measure is a sum across five weeks so missing entries will strongly affect the magnitude of the calculated outcome sum. For RQ4 and RQ5, no imputation will be applied as these analyses result in

¹¹ We understand that this guidance will be made available during the summer of 2024. This date is after planned publication of the study plan so the recommendations are not included here, however it will be before data analysis takes place so can be implemented for this additional analysis.

¹² The diary is designed so that weeks cannot be skipped and weekly entries must be completed in order, within two blocks of five. Therefore any imputed missing entries would only be from weeks 4 and 5 and/or weeks 9 and 10.

proportions. In all analyses, respondents with fewer than three weeks of responses in a time period will not be included in the analysis for that time period.

Rate of missing data will be reported for each week of the study in each of the ChatGPT and Non-GenAI groups. We will run a chi-squared test to explore whether missing diary entries are unevenly split across the ChatGPT and Non-GenAI groups. Additionally, since timing of dropout from the trial would be considered informative, a survival analysis will be undertaken modelling the time to the first missing diary entry for each teacher and comparing the hazard between the ChatGPT and Non-GenAI groups. This analysis would inform us if dropout from the trial occurs at a similar or different rate among the trial groups. If there is no evidence that missing data is occurring unevenly between the groups, multilevel models are considered robust to missing data. If there is evidence of unevenness, our assumption that data are missing at random is likely not sound, so a censored regression analysis will be implemented as a sensitivity check such as a Tobit model. This model would assume that data are missing because workload is too high and therefore outcome variables are right-censored.

In addition, as sensitivity analysis, we will repeat the primary analysis for the subset of teachers who have completed all five weeks of diary entry (w6-10) i.e. a complete case analysis.

We anticipate that missing lesson resources are unlikely to be random in that teachers will probably submit what they consider to be their best lesson resources and any missing resources (e.g. if they only submit one of the three requested) would likely have been of lower quality than the ones they submit. However, it remains possible that the rate at which lesson resources are missing is independent of randomisation group status, which would make this less of a challenge to internal validity. We will run a chi-squared test to explore whether missing lesson plans are unevenly split across the ChatGPT and Non-GenAI groups. If the split seems even, we will deal with the missing lesson resources by sampling only from teachers who have submitted resources for all three lessons that we asked for. If there is evidence that the rate of missingness differs significantly between the groups, we will undertake additional analyses to quantify the potential level of bias introduced. To achieve this, we will ensure that missing lessons can be included in the sampling. Where a missing lesson resource is sampled, we will also sample an additional resource chosen at random from the same arm, so that the panel is guaranteed to review 40 resources. We will randomly assign the missing resource a rank of between 30 and 40 (the bottom 25% of the range - higher numbers denote lower ranks). We will then conduct two Mann Whitney U tests one comparing the ranks of the resources with the missing resources replaced with other lesson resources and one where the missing resources have been randomly ranked in the bottom quartile. The scenario where missing resources are replaced with a randomly selected alternative illustrates the assumption that resources are missing at random whereas the scenario where missing resources are given a low rank illustrates the assumption that missing resources are more likely to be of low quality. We will report the results of both tests and use any differences to inform our commentary about the effect of missing data in our analysis of lesson resources.

Implementation and process evaluation (IPE) design

Research questions

The aim of the IPE will be to add context to the impact analyses described above. Key areas of teacher outcomes include their approaches to lesson and resource preparation, their confidence in using GenAI/ChatGPT, and their confidence in lesson and resource preparation more generally. Perceived pupil outcomes (including for disadvantaged pupils) include engagement and learning in science. The IPE will also explore the role of the ChatGPT teacher guide as a mechanism for promoting effective ChatGPT use.

The IPE research questions are shown below, along with the IPE dimensions covered.

RQ6 To what extent do teachers adhere to their allocated approaches? Adherence

- To what extent do teachers allocated to the ChatGPT approach use ChatGPT for lesson and resource preparation?
- Do teachers allocated to the Non-GenAI approach adhere to their condition?
- What are the reasons for adherence or non-adherence?

This research question will contextualise the potential compliance analysis within our impact evaluation.

RQ7 How is ChatGPT used by science teachers while preparing for lessons? Fidelity of implementation, adaptation

- Which activities (e.g., use cases identified in the teacher guide) do teachers use it for and why? Does this vary between weeks 1-5 (the learning period) and weeks 6-10?
- What are teachers' perceptions of the process of using ChatGPT?
- What are the facilitators and barriers to using ChatGPT for lesson/resource preparation?
- What are the perceived benefits, drawbacks and risks of using ChatGPT for these activities? How do teachers respond to these?
- How does the approach to lesson preparation compare in schools using ChatGPT to those not using GenAI?

This research question will help understand how, why and in what circumstances teachers use ChatGPT, adding context to the impact evaluation.

RQ8 How do teachers use the teacher guide (website)? Teacher guide

- How do teachers use the teacher guide across the trial period (e.g. frequency, different use cases)?
- How accessible, useful and relevant do teachers think the ChatGPT teacher guide is?
- Are teachers aware of the principles and recommended practice outlined in the guide?
- To what extent does teachers' use of ChatGPT (RQ6) align with the principles and recommended practice outlined in the guide?
- How long did teachers perceive they needed before they could use ChatGPT effectively?

This research question complements the measure of frequency of using the teacher guide (RQ5) within our impact evaluation.

RQ9 What is the perceived impact of using ChatGPT in lesson and resource preparation? Perceived impact, unintended consequences, cost evaluation

- How do teachers perceive the quality of ChatGPT output compared to other sources of science teaching resources?
- How do pupils respond to lessons/resources prepared using ChatGPT compared to those prepared without using GenAI? Is there any difference in response for FSM pupils?
- If time is saved when using ChatGPT, what is this time used for? (does it represent a reduction in workload, or is time diverted elsewhere?)
- If time spent/workload remains the same or increases when using ChatGPT, what do teachers think are the reasons for this?
- Were there any direct financial costs/savings for teachers and schools when using ChatGPT?
- What are teachers' intentions for using ChatGPT/GenAI after the trial period?

This research question will add context to any change in workload associated with using ChatGPT (RQ1) by exploring teachers' perceptions of and reasons for any changes in time use, and comparing this with teachers' self-reported changes in workload. It will add context to the quality comparison of lessons/resources planned using ChatGPT and Non-GenAI approaches (RQ3) within our impact evaluation by considering teachers' perceptions of lesson quality, and by comparing teacher-level perceptions of quality with the lesson plan panel scores. In case study schools, we will also gather teacher and pupil perspectives on lesson quality for ChatGPT and Non-GenAI arms, to complement the panel quality comparison.

RQ10 To what extent do moderators affect behaviour and workload changes? Moderators

- How does teacher confidence in science content, science pedagogy and technology use (1) affect the proportion of lessons for which ChatGPT is used and (2) moderate changes in teacher workload?
- Do teachers use ChatGPT differently when planning for a topic they have not taught before?
- To what extent does shared/centralised lesson planning interact with use (or not) of GenAI?

RQ11 What is usual practice in science teachers' lesson and resource preparation, and use of GenAI? Usual practice

- How do science teachers usually prepare lessons and resources?
- How does this vary across schools?
- For Non-GenAI teachers, how has implementing a Non-GenAI approach changed their preparation, if at all?
- How did science teachers use GenAI prior to the trial?

The IPE design is aligned to the trial logic model (Figure 1), including:

- Teacher Inputs: exploring teacher use and perceptions of ChatGPT guide
- Teacher Outputs: describing how and why teachers use ChatGPT during lesson and resource preparation

- Teacher Outcomes/Impacts: exploring perceived outcomes of the two approaches, changes in teacher confidence, and how any saved time is used
- Pupil Outcomes/Impacts: describing perceived pupil engagement and learning
- Moderators: testing the moderating effects of (1) teacher confidence and (2) the extent of shared/centralised planning

From previous research, it is not clear whether to expect a differential impact for FSM pupils, especially as the evidence base on use of GenAI in schools at a very early stage. At school-level, although most teachers and leaders (72%) disagree that their workload is acceptable, this proportion is higher in low-FSM schools compared with high-FSM schools (Schmidt *et al.*, 2009) (Adams *et al.*, 2023), suggesting that workload reduction may particularly benefit low-FSM schools. We will explore teachers' perceptions of any differential impact for FSM pupils, and the mechanisms for this, through the case study visits.

Research methods

The IPE research methods are summarised in **Error! Reference source not found.5** and described below.

Table 5: IPE methods overview

			IPE dimension/ research area						Data analysis methods
			Adherence	Fidelity, adaptation	Teacher guide	Perceived impact, unintended consequences	Moderators	Usual practice	
			RQ6	RQ7	RQ8	RQ9	RQ10	RQ11	
Data collection approaches and sample	Diary	All teachers	X	X			X		Descriptive statistics
	Surveys	All ChatGPT teachers	X	X	X	X	X	X	Multi-level modelling (moderator analysis)
	Quiz		X	X	X				Factor analysis (confidence measures)
	ChatGPT transcripts	Random sample 20 lessons		X	X				Descriptive statistics
	Senior leader interview	12 case study schools Stratified sampling				X		X	Content analysis
	Teacher focus group		X	X	X	X		X	Thematic analysis
	Preparation walkthrough		X	X	X	X			
	Pupil focus group					X			

Diary

As outlined in the impact section, all teachers will be asked to complete a weekly diary which captures their workload for lesson and resource preparation, any use of GenAI tools by Non-GenAI teachers (to assess contamination) and use of ChatGPT and the guide by ChatGPT teachers.

Teacher Surveys

All participating teachers will be asked to complete surveys at baseline (Feb-Mar 2024) and endpoint (July 2024). The baseline survey covered teacher background characteristics (teaching experience and specialism, including whether they have previously taught the Y7/8 science topics included in the trial period), teacher practice (including usual lesson preparation and workload) and teacher confidence (lesson preparation, science teaching and technology/GenAI use, incorporating the Technology, Pedagogy, and Content Knowledge (TPACK) model (Schmid, Brianza and Petko, 2020)).

The endpoint survey will repeat key questions from the baseline survey (teacher practice and teacher confidence) to monitor changes. It will also include additional questions for teachers in the ChatGPT group about ChatGPT use (use types, benefits/drawbacks, facilitators/barriers) and perceived pupil engagement.

Quiz

As outlined in the impact section, all teachers in the ChatGPT arm will be asked to complete a quiz to check teachers' awareness of the principles and recommended practice set out in the teacher guide.

ChatGPT transcripts

All teachers in the ChatGPT group will be asked to submit lesson resources for three lessons, along with any ChatGPT transcripts used to generate them. Twenty ChatGPT lessons and associated transcripts will be randomly selected for analysis, subject to capping at one lesson per school. Our impact evaluation (RQ3) will assess the sampled lesson resources while our IPE evaluation will assess the sampled ChatGPT transcripts. Analysis of these transcripts is described in the Analysis section below.

School case studies

We will undertake 12 face-to-face case studies of schools from both groups – ChatGPT (n=8) and Non-GenAI (n=4). We have included more schools from the ChatGPT arm, in order to stratify the sample by high/low engagement with ChatGPT (from analysis of early teacher diaries). For the Non-GenAI arm, if possible, we will stratify by previous experience of using GenAI, to include schools where this approach is 'business as usual' as well as schools where the Non-GenAI approach constrains their practice. Where possible, we will aim to include schools across a range of characteristics (Ofsted rating, school type, different planning approaches).

Case study visits will take place in June-July 2024. The aim of the case studies is to gain an in-depth understanding of the use of ChatGPT for lesson and resource preparation and a comparison to usual practice.

We intend to study a relatively high number of cases (12/58 schools) because we expect high heterogeneity in teachers' use of ChatGPT and its impact, which we want to capture. The limited extant

research on the use and impact of GenAI in lesson preparation has taken a broad descriptive approach which neglects heterogeneity and its causes, while we conceptualise GenAI use and impact in relation to existing practice and culture of lesson planning (at department and teacher level) and teacher experience and capabilities in science teaching and using GenAI. Analysing 8 ChatGPT cases will enable us to characterise ChatGPT use and outcomes in relation to these characteristics. Describing the department and teacher contexts for each case will also enhance recognisability by supporting teachers and leaders to identify complete cases, or elements of cases, which are similar to their own contexts. Our approach to analysing and comparing cases is described further in the Analysis section below. In addition, a case study methodology will allow us to triangulate teacher and pupil perceptions of pupil outcomes, and contextualise these in terms of previous practice in lesson planning, and (for ChatGPT teachers) the specific ways that ChatGPT has been used for their lessons. This exploratory work will provide insights for any future work on pupil outcomes.

The 12 case studies will aim to include¹³:

- A curriculum leader (e.g. Head of Science) interview, focusing on usual practice in lesson preparation, use of AI, and workload, and (for ChatGPT schools) their experience and perceived impact of teachers using ChatGPT for lesson and resource preparation.
- an individual/group interview with up to 5 science teachers. For schools with ≤ 5 teachers participating, all teachers will be invited. For schools with >5 teachers, we will ask the trial lead to include teachers across the departmental range of teaching experience and engagement with the trial. The interview will focus on their usual practice of lesson preparation and (for ChatGPT teachers) how they use ChatGPT and perceived benefits/drawbacks and impact. For Non-GenAI schools who have previously used GenAI to support lesson and resource preparation, we will gather information on their previous use of GenAI.
- a lesson preparation walkthrough (1 science teacher), focusing on their approach to lesson preparation, e.g., what activities they undertake and why, and how they adapt lessons and resources for the specific class. For ChatGPT teachers, this will also include how they use ChatGPT. For Non-GenAI teachers, this will also include whether/how their usual practice has changed while being asked not to use GenAI.
- a focus group of 3-5 pupils from one Y7 or Y8 class, focusing on their experience of science lessons, and perceived engagement/learning. Where possible, these pupils will be taught by the teacher from the lesson preparation walkthrough, to provide pupils' perspectives on the approach to lesson preparation.

Although we anticipate the main cost to be related to teachers' time (captured through the impact analysis), we will also gather data on any other perceived costs through the interviews.

Analysis

School case studies

Qualitative data from the school case studies (n=12) will include senior leader interviews, teacher focus groups, lesson and resource preparation walkthroughs, and pupil focus groups. This data will be recorded,

¹³ Some schools signed up to the trial with only one or two teachers. We will be mindful of number of participating teachers per school but we note that we may need to be flexible about the composition of the case study activities if the number of participating teachers in a case study school is small.

transcribed, and analysed thematically in MAXQDA. We will develop a deductive top-level coding frame based on the research questions (e.g. fidelity and adaptations, teacher guide), and will code the data from each source inductively within that frame. Using a deductive frame for top-level coding ensures that coding is focused on the research questions, while inductive coding within that frame ensures that all data relevant to the research questions is captured. The inductive coding will then be compared with the logic model, including developing a context-specific description of the COM-B model (Michie, van Stralen and West, 2011), mapping how teachers described their capability, opportunity and motivation to use ChatGPT, and their subsequent use of ChatGPT. This comparison is completed after inductive coding to minimise selective coding based on the expected outputs and outcomes from the logic model.

Case-oriented thematic analysis (Miles, Huberman and Saldaña, 2019) will provide a rich description of implementation for each school case (n=12), which will include a description of key school and teacher characteristics (e.g. extent of shared planning, teacher confidence (as described in the analyses below), extent of using ChatGPT, and prior use of GenAI) from the quantitative survey and diary data. We will create a summary of each case, such as a matrix or network display.

We will establish an audit trail by keeping copies of field notes, transcribed data, coded data, and case summaries, which will be available to colleagues involved in quality assurance of the IPE. At least two experienced qualitative researchers will contribute to analysis to enable peer discussion of findings, and key methodological discussions and decisions will be noted and reported. To support trustworthiness, we will include our coding frame in the trial report. We will also share the IPE analysis process and emerging findings with the study advisory board, as a form of peer debriefing.

ChatGPT transcripts

For each sampled transcript, we will categorise the activity ChatGPT was used for (one of the six use cases in the teacher guide, or any other activity), and describe any additional activities beyond the six use cases. We will assess the transcript against a brief yes/no checklist of the practice presented in the teacher guide, e.g. (1) providing relevant teaching context, (2) refining prompts (3) avoiding personal data. We will report the proportion of transcripts which align with each point in the checklist.

In addition, where 'quality and accuracy of scientific content' in a lesson is rated poorly in the impact evaluation, we will check whether the same errors are present in the ChatGPT transcript.

Quantitative data – teacher diary

We will summarise the diary data with descriptive statistics, including frequencies and distribution statistics (e.g. mean, median, quartiles and min/max) to describe teacher preparation workload in each trial arm, contamination, and use of ChatGPT and the guide.

Quantitative data – teacher surveys

We will summarise the data from baseline and endpoint surveys with descriptive statistics, including frequencies and selected cross-tabulations. To test and elaborate the logic model, we will describe teachers' opportunity, capability and motivation to use GenAI for lesson preparation, the activities they use ChatGPT for, their perceptions of resource quality, and how they use any time saved. We will compare baseline and endpoint teacher data to explore changes in individual teacher practice and confidence, including comparing changes for the two trial arms.

Quantitative data – quiz (ChatGPT group only)

We will describe the score distribution to assess teachers' overall awareness of the ChatGPT guide content, and response frequencies for each item to describe teachers' awareness of specific principles/practice and any common misconceptions. We will cross-tabulate overall scores with high and low ChatGPT use. Triangulated with the teacher perceptions reported in the surveys, this will help us to assess teachers' capability to use ChatGPT, and how this relates to ChatGPT use.

Quantitative data – teacher confidence, use of ChatGPT, and workload changes (ChatGPT group only)

We will look at the relationship between use of ChatGPT, teacher confidence and changes in workload. We will first conduct a polychoric factor analysis on the items in the baseline teacher survey relating to confidence¹⁴. Once we have established the factor structure, we will use the factors to answer the following three questions:

- 1) How is teacher confidence associated with the proportion of lessons for which ChatGPT is used? (ChatGPT group only)
- 2) How does teacher confidence moderate changes in teacher workload? (both randomisation groups)
- 3) To see if there is evidence that changes in teacher confidence are influenced by use of ChatGPT. (ChatGPT group only)

More information about the analysis for each question is shown below.

1) How is teacher confidence associated with the proportion of lessons for which ChatGPT is used?

To answer this question, we will build a binomial multilevel model with a logit link function to model how the proportion of lessons for which ChatGPT is used in the initial five-week period is affected by baseline teacher confidence factors. The model will account for the clustering of teachers in schools by modelling it as a random effect. The regression model is defined by:

$$\text{logit}(p\text{ChatGPT}_{ij}) = \sum_{f=1}^N \beta_f \text{factorscore}_f + \beta_{0j} + \beta_2 \text{BaseAIuse}_{ij}$$

Where:

- $p\text{ChatGPT}_{ij}$ is the probability that teacher i in school j uses ChatGPT to aid lesson preparation. This will be measured using two questions from the weekly teacher diary and parameterised as the total number of 'successes' and 'failures' across the five weeks for each teacher.
- BaseAIuse_{ij} is a measure of how much teacher i in school j used generative AI prior to the study as measured by the teacher baseline survey.
- β_{0j} is the intercept in school j (modelled as a random effect).

¹⁴ Including the items adapted from TPACK

- β_f is the coefficient of interest estimating the association between factor f and the proportion of lessons planned using ChatGPT.
- N is the number of factors (or different dimensions) identified in the previous factor analysis.
- β_2 is a coefficient estimating the association between baseline teacher AI use and the proportion of lessons planned using ChatGPT.

2) How does teacher confidence moderate changes in teacher workload?

We will build a series of linear multilevel models (one for each factor) similar to that used in the primary analysis but with the addition of main effect and interaction terms to measure moderating effect of each factor, controlling for baseline lesson planning and resource preparation time by means of a covariate and accounting for clustering of teachers at school level by modelling it as a random effect. Each teacher-level regression model is defined by:

$$\text{Workload}_{ij} = \beta_{0j} + \beta_1 \text{ChatGPT}_j + \beta_2 \text{BaseWorkload}_{ij} + \beta_3 \text{factorscore}_{ij} + \beta_4 \text{factorscore}_{ij} * \text{ChatGPT}_j + \varepsilon_{ij}$$

Where:

- Workload_{ij} is the preparation workload of teacher i in school j for years 7 & 8 summed across the second five-week period. This will be measured using a weekly teacher diary.
- BaseWorkload_{ij} is the estimated weekly preparation workload for Year 7 and 8 science lessons for teacher i in school j as measured by the baseline teacher survey.
- β_{0j} is the intercept in school j (modelled as a random effect).
- β_1 is a coefficient of interest estimating how much difference being assigned to the ChatGPT group makes to teacher workload.
- β_2 is a coefficient estimating the association between baseline teacher workload affects workload in the second five-week period.
- ChatGPT_j is a variable indicating whether school j was randomly assigned to use ChatGPT or not.
- factorscore_{ij} is the factor score for teacher i in school j.
- β_3 is a coefficient of interest estimating the association between the confidence factor and teacher workload as a main effect.
- β_4 is a coefficient of interest estimating how the interaction between the confidence factor score and group assignment affects teacher workload.
- ε_{ij} is the residual error term for teacher i in school j.

To avoid the scenario where teachers who only teach one or two lessons unduly influence the results, we will weight the model by the total number of lessons taught to qualifying year-groups in the time period.

3) *How does use of ChatGPT affect teacher confidence?*

For this analysis we will first calculate the confidence factor scores for each teacher at endpoint by projecting the factor loadings calculated from the baseline survey onto the same items in the endpoint survey. For each factor we will conduct a paired t-test on the difference between the scores at baseline and endpoint (for the ChatGPT group only) to establish which confidence factors, if any, show evidence of having changed during the study. For each factor where there is a significant change, we will construct a linear multilevel model to explore how the endpoint factor score is affected by the proportion of lessons in the last five weeks for which the teachers used ChatGPT for their lesson preparation controlling for baseline factor score by means of a covariate and accounting for clustering of teachers at school level by modelling it as a random effect. The teacher-level regression model is defined by:

$$\text{EndlineFactor}_{ij} = \beta_{0j} + \beta_1 \text{PropChatGPT}_{ij} + \beta_2 \text{BaseFactor}_{ij} + \varepsilon_{ij}$$

Where:

- **EndlineFactor_{ij}** is the confidence factor score of teacher i in school j at the end of the study. This will be measured using the endpoint survey.
- **β_{0j}** is the intercept in school j (modelled as a random effect).
- **β₁** is a coefficient of interest estimating the association between the proportion of lessons that use ChatGPT and the endpoint factor score.
- **β₂** is a coefficient estimating the association between the baseline and endpoint factor scores.
- **PropChatGPT_{ij}** is the proportion of lessons prepared with the help of ChatGPT for teacher i in school j.
- **BaseFactor_{ij}** is the baseline factor score of teacher i in school j.
- **ε_{ij}** is the residual error term for teacher i in school j.

To avoid the scenario where teachers who only teach one or two lessons unduly influence the results, we will weight the model by the total number of lessons taught to qualifying year-groups in the time period.

Synthesis of IPE data

We will collate and triangulate data sources to provide a rounded picture of how both preparation planning approaches are enacted, and variation across schools and teachers. This will include identifying any patterns based on key moderators. The IPE findings will contextualise impact findings and aid their interpretation. We will use the data from the IPE, along with the impact findings, to update the theory of change at the end of the project.

Ethics and registration

The trial will be designed, conducted and reported to CONSORT standards (<http://www.consort-statement.org/consort.statement/>) and registered on <http://www.controlled-trials.com/>.

This evaluation will be conducted in accordance with NFER's Code of Practice, available at [NFER Code of Practice](#). All of NFER's projects abide by its Code of Practice, which is in line with the Codes of Practice from BERA (the British Educational Research Association), MRA (the 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. NFER and EEF will work together to also ensure each organisation's policies can be applied in practice.

Agreement for participation within the trial was provided by the headteacher via signing the MoU that outlines the responsibilities of all parties involved in the trial.

A separate opt-out consent process will be used for the pupil focus groups and will only apply to those selected to participate. We will provide schools with information to share with parents in advance of the case study visit. Parents/carers will be given a written information sheet about the focus groups which will contain full details about the focus group and what their child will be asked to do. If the parent/carer does not wish for their child to participate then they should complete and return the form to the school in advance of the visit. The school will collate this information and pupil personal data will not be processed by the research team

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

Data protection

All data gathered during the trial will be held in accordance with the Data Protection Act 2018 and General Data Protection Regulation (GDPR) and will be treated in the strictest confidence by NFER and EEF. No school or teacher will be named in any report arising from this work, nor will we include any information that might mean that someone else could identify them.

NFER is the data controller for this evaluation and makes decisions about what personal data is used and how it is processed in accordance with the objectives of the evaluation set by the EEF. After the report is published, teacher responses from the diaries, quiz and teacher surveys may be transferred and stored in the EEF archive for future research, at which point the EEF will become the data controller for the archived data.

The legal basis for processing personal data is covered by GDPR Article 6 (1) (f): Legitimate interests: the processing is necessary for your (or a third party's) legitimate interests unless there is a good reason to

protect the individual's personal data which overrides those legitimate interests. A legitimate interest assessment has been undertaken. The trial fulfils one of NFER's core business purposes (undertaking research, evaluation, and information activities). It has broader societal benefits and will contribute to improving the lives of learners by providing evidence about the impact of teaching techniques used in the classroom. Research cannot be done without processing personal data, but processing does not override the data subject's interests.

NFER has provided an MoU to schools, explaining the nature of the data being requested of schools and teachers, how it will be collected and processed. The privacy notice for this trial is available at https://www.nfer.ac.uk/media/wntddbxxo/eeai_school_teacher_privacy_notice.pdf.

As part of the sign-up process, NFER collected Expressions of Interest (EOI), which included the names, contact details, role of the individual completing the form, their school and how they heard about the project. Upon receiving the EOI, NFER requested a completion of the MoU from the headteacher or a member of the senior management team. The MoU asked for the names, contact details and job role for the headteacher and the trial lead.

Further personal data about teachers will be collected throughout the trial via teacher data template, weekly diary entries, online surveys, interviews and case study focus groups. NFER will collect teachers' personal data through these activities.

NFER will use Questback to provide online surveys. See https://www.questback.com/assets/uploads/Survey_Privacy_Policy.pdf for further information.

Microsoft Teams or Zoom may be used for interviews which cannot be undertaken in person. Privacy notices for both online communication tools are available:

Microsoft Teams – <https://docs.microsoft.com/en-us/microsoftteams/teams-privacy>

Zoom - <https://explore.zoom.us/en/privacy/>

Personal data collected through these activities include names, contact details, job role, length of time teaching, subject specialism, prior use and confidence in GenAI, confidence in science lesson preparations, their knowledge, skills, confidence and attitudes to technology and pedagogical beliefs, perceptions of facilitators, barriers and challenges as well as perceived costs in using GenAI and teacher perceptions of pupil experiences.

Data archiving and deletion

After the report is published (currently planned for December 2024), teacher responses from the diaries, quiz and teacher surveys may be transferred and stored in the EEF archive for future research. The EEF archive is managed by FFT on behalf of EEF and hosted by the Office of National Statistics (ONS). The teacher data may also be shared with the Department for Education (DfE) and linked with information about teachers from the ONS. Names and other meaningful identifiers are removed before the data is added to the EEF archive. At this point EEF becomes the data controller for the teacher data. The EEF will keep information in the EEF archive for as long as it is needed for research purposes. Please see EEF's archive guidance [here](#) and EEF's privacy notice [here](#) for more information on how EEF processes and will use your data. EEF and other research teams will be able to access the data as part of subsequent research through the ONS Approved Researcher Scheme. The Approved Researcher Scheme is used by the ONS to grant

secure access to data that cannot be published openly, for statistical research purposes, as permitted by the Statistics and Registration Service Act 2007 (SRSA).

Audio recordings from teacher interviews will be transcribed and deleted within one month of the interview date. All other personal data held by NFER will be deleted within one year of publication of the final report, currently expected to be December 2025.

Personnel

Table 6: Key members of the study team

Name	Organisation	Role and Responsibilities
Helen Poet	NFER	Project Director – responsible for overall delivery of the trial
Palak Roy	NFER	Trial Manager – day-to-day management of the trial, delivery of the trial design and impact evaluation lead
Katherine Aston	NFER	IPE Lead – design and delivery of the IPE
Ruth Staunton	NFER	Trial statistician – lead quantitative analysis for the main trial
David Thomas	NFER	Science assessment lead – oversee RQ3 on lesson resource quality
Daniel Jackson	NFER	IPE researcher – undertake IPE case studies
Kathryn Hurd	NFER	Research Operations Lead - overall data collection and setting communications strategy
Jo Stringer	NFER	Senior Project & Delivery Manager– day-to-day operations for the main trial including coordinating evaluation data collection and point of contact for the settings
Lydia Wallis	NFER	Project and Delivery Manager – data collection and setting communications for the Formative Evaluation
Faizaan Sami	EEF	Evaluation Manager
Amy Ellis-Thompson	EEF	Senior Programme Manager
Christine Kelly	EEF	Methodological Innovation Lead
James Turner	Hg Foundation	CEO Co-funder for the evaluation
Tim Harrison	Hg Foundation	Data Lead Co-funder for the evaluation

Sarah McMorris	Bain and Company	Senior Manager Responsible for the teacher guide
Olivia Wilkinson	Bain and Company	Senior Consultant Responsible for the teacher guide

Study Advisory Board

Name	Organisation	SAB Specialism
Sam Sims	Lecturer, Institute of Education, UCL; Research Lead, Ambition Institute	Evaluation expertise, workload subject expertise
Manolis Mavrikis	Professor of Artificial Intelligence and Analytics in Education, UCL	Evaluation expertise; edtech subject knowledge
Chris Goodall	Head of Digital Education, Bourne Education Trust	Using GenAI in teaching expert
Aditi Bhutoria	Assistant Professor, Indian Institute of Management, Calcutta; OpenDevED collaborator	Methodological expertise; Edtech subject expertise; practitioner experience
Rachel Dulley	KS4/KS5 Science teacher and lead practitioner for teaching and learning, Ravensbourne School	Science teaching experience
Bernadette Delahunty	KS3-KS5 Science teacher, Greensward Academy	Science teaching experience

Risks

Table 7: Evaluation risk assessment

Risk	Assessment	Controls, countermeasures, and contingencies
Tight timetable for recruitment and baseline data collection results in insufficient schools and teachers recruited to the trial	Likelihood: low Impact: high	<ul style="list-style-type: none"> NFER's research operations team will monitor recruitment. If required, decide and monitor pre-agreed recruitment targets to identify any unfavourable trends early on to act quickly. Discuss the possibility of drawing a top-up sample; communicate with early EOI schools and speak to EEF to promote the trial on their social media more frequently.
Teacher Choices are not well implemented or	Likelihood: low to moderate Impact: high	<ul style="list-style-type: none"> Headteacher signs MoU with clear identification of requirements. Clear initial and ongoing communications from NFER sent directly to the participants that includes

Non-GenAI group adopts aspects of ChatGPT which leads to contamination		participation expectation. In addition to this, NFER sends the group allocation and a set of guidelines for implementation that includes Do's and Don't's directly to the teachers.
Schools and/or teachers do not complete the data on implementation fidelity	Likelihood: moderate Impact: moderate	<ul style="list-style-type: none"> • Clear initial and ongoing communication with participants explaining trial expectations. • Incentive payments to all schools are also attached to implementation data.
Schools and/or teachers drop out from trial and primary analysis.	Likelihood: moderate Impact: high	<ul style="list-style-type: none"> • Clear initial and ongoing communication with participants explaining principles and expectations. Sign up to the trial via Memorandum of Understanding with clear identification of requirements. NFER to communicate with one key contact per school (the trial lead) to inform them of next steps. • Where possible, reduce data collection burden and in case of lower response rate to multiple instruments, request participants to respond to the most important instrument. • Over-recruit schools by assuming 10% attrition from primary outcome. • Incentive payments to all schools upon completion of evaluation activities. Consider break down of incentive payments to encourage a higher response during weeks 6-10 even if teachers did not respond to all weeks 1-5 diaries.
School lead (for the trial) attrition (e.g., school's lead for the trial becomes unresponsive or leaves the school during the course of the trial)	Likelihood: low Impact: low	<ul style="list-style-type: none"> • Establish contact with the headteacher who signed the MOU. • NFER to request a different key contact person if the school is amenable.
Participants upload confidential information and/or intellectual property on ChatGPT	Likelihood: low Impact: high	<ul style="list-style-type: none"> • NFER will provide trial guidance to all teachers in the ChatGPT group. This will explicitly mention not to enter any personal data in ChatGPT. Via this guidance, teachers are also encouraged to read DfE's guidance in protecting data for pupils and staff and to adhere to their school's policies whilst using ChatGPT

Timeline

All trial activities are undertaken by NFER unless otherwise specified below.

Table 8: Timeline of activities for the trial

Dates	Activity (organisations responsible/leading)
December 2023 – January 2024	<ul style="list-style-type: none"> IDEA workshop and project set-up meetings (NFER, EEF, Hg Foundation and Bain and Company) Complete project set-up (including due diligence and data protection impact assessment) Finalise recruitment documents (NFER and EEF)
February – March 2024	<ul style="list-style-type: none"> Promote the trial on website and social media (NFER and EEF) School recruitment via EoI and MoUs Teacher data collection Baseline Teacher survey Draft the Quiz questions to use in compliance (Bain and Company) Randomisation and inform participants about their group allocation Draft Study Plan/SAP
March – April 2024	<ul style="list-style-type: none"> Finalise online diary and quiz First Study Advisory Board meeting
April– May 2024	<ul style="list-style-type: none"> Teachers adopt allocated ‘choice’ to the first block of five weeks Teachers complete the diary every week ChatGPT teachers complete the teacher guide quiz Update SAP if required post-randomisation Finalise Study Plan
June – July 2024	<ul style="list-style-type: none"> Study plan published on EEF website Trial registration completed on ISRCTN Teachers adopt allocated ‘choice’ to the second block of five weeks Teachers complete the diary every week NFER undertakes school case study visits Teachers provide lesson resources Teachers complete endpoint survey
August – October 2024	<ul style="list-style-type: none"> Data processing, analysis and report writing First draft trial report submitted to EEF Study Advisory Board meeting
November – December 2024	<ul style="list-style-type: none"> Trial report revisions (NFER, EEF, Hg Foundation and Bain and Company) Draft and finalise Teacher accessible output Teacher accessible output and trial report published on EEF website

References

- Adams, L., Coburn-Crane, S., Sanders-Earley, A., Harris, H., Taylor, J. and Taylor, B. (2023) *Working lives of teachers and leaders – Wave 1*. Available at: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1148571/Working_lives_of_teachers_and_leaders_-_wave_1_-_core_report.pdf (Accessed: 6 September 2023).
- Bulus, M., Dong, N., Kelcey, B. and Spybrook, J. (2021) 'PowerUpR: power analysis tools for multilevel randomized experiments 1.1.0'. Available at: <https://cran.r-project.org/web/packages/PowerUpR/index.html> (Accessed: 21 November 2023).
- ChatGPT (2024) *OpenAI*. Available at: <https://openai.com/chatgpt/> (Accessed: 13 June 2024).
- Department for Education (2023) *Generative AI in education. Call for evidence: summary of responses*. Available at: https://assets.publishing.service.gov.uk/media/65609be50c7ec8000d95bddd/Generative_AI_call_for_evidence_summary_of_responses.pdf (Accessed: 20 March 2024).
- Education Endowment Fund F (2022) *Statistical analysis guidance for EEF evaluations*. Available at: <https://d2tic4wvo1iusb.cloudfront.net/production/documents/evaluation/evaluation-design/EEF-Analysis-Guidance-Website-Version-2022.14.11.pdf?v=1709155078> (Accessed: 6 March 2024).
- Fletcher-Wood, H. (2023) 'How to improve behaviour and wellbeing, and how you're using AI in schools', *Teacher Tapp*, 28 November. Available at: <https://teachertapp.co.uk/articles/how-to-improve-behaviour-wellbeing-and-how-youre-using-ai-in-schools/> (Accessed: 18 June 2024).
- Keegan, G. (2023) 'Education Secretary addresses BETT 2023'. *BETT 2023*, ExCel London, 29 March. Available at: <https://www.gov.uk/government/speeches/education-secretary-addresses-bett-2023> (Accessed: 12 June 2024).
- Malik, A. (2023) 'OpenAI's ChatGPT now has 100 million weekly active users', *TechCrunch*, 6 November. Available at: <https://techcrunch.com/2023/11/06/openais-chatgpt-now-has-100-million-weekly-active-users/> (Accessed: 12 June 2024).
- Michie, S., van Stralen, M.M. and West, R. (2011) 'The behaviour change wheel: a new method for characterising and designing behaviour change interventions', *Implementation Science*, 6(42), pp. 1–11. Available at: <https://doi.org/10.1186/1748-5908-6-42>.
- Miles, M.B., Huberman, A.M. and Saldaña, J. (2019) *Qualitative data analysis: a methods sourcebook*. 4th edition. London: Sage.
- Organisation for Economic Co-operation and Development (2018) 'TALIS 2018 database'. Available at: <https://www.oecd.org/education/talis/talis-2018-data.htm> (Accessed: 12 June 2024).
- Organisation for Economic Co-operation and Development (2019) 'TALIS 2018 Results (Volume I). Teachers and school leaders as lifelong learners'. Paris: OECD Publishing.
- Schmid, M., Brianza, E. and Petko, D. (2020) 'Developing a short assessment instrument for Technological Pedagogical Content Knowledge (TPACK.xs) and comparing the factor structure of an integrative and a transformative model', *Computers & Education*, 157, p. 103967. Available at: <https://doi.org/10.1016/j.compedu.2020.103967>.

Schmidt, D.A., Baran, E., Thompson, A.D., Koehler, M.J., Mishra, P. and Shin, T. (2009) *Survey of preservice teachers' knowledge of teaching and technology*. Available at: https://news.cehd.umn.edu/wp-content/uploads/2009/06/tpck_survey.pdf (Accessed: 13 June 2024).

Tabachnick, B.G. and Fidell, L.S. (2013) *Using multivariate statistics*. 6th Edition. Boston: Pearson.

United Nations Educational, Scientific and Cultural Organization (2023) *Guidance for generative AI in education and research*. Available at: <https://unesdoc.unesco.org/ark:/48223/pf0000386693/PDF/386693eng.pdf.multi> (Accessed: 12 June 2024).

Walker, M., Worth, J. and Van den Brande, J. (2019) *Teacher workload survey 2019: technical report*. Available at: https://assets.publishing.service.gov.uk/media/5e12fcc9e5274a0fa4dc2894/teacher_workload_survey_2019_technical_report__amended.pdf#page=16 (Accessed: 26 March 2024).

Whittaker, F. (2023) 'ChatGPT: one in three teachers use AI to help with school work', *Schools Week*, 14 September. Available at: <https://schoolsweek.co.uk/chatgpt-one-in-three-teachers-use-ai-to-help-with-school-work/> (Accessed: 12 June 2024).

Appendix A

Table A1: Overview of research objectives

Research questions	Methodology area	Data collection methods & participants	Data analysis methods
RQ1 (Primary): What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons over five weeks, after five weeks of initial use?	Impact	Teacher weekly diary (weeks 6–10)	Linear multilevel models
RQ2: What is the impact of using ChatGPT on teacher lesson and resource preparation time for Year 7 & 8 science lessons during the initial five week learning period?	Impact	Teacher weekly diary(weeks 1–5)	
RQ3: What is the effect of using ChatGPT on the quality of lesson and resource materials used in Year 7 & 8 science lessons?	Impact	Teacher lesson resources (weeks 6–10)	Mann-Whitney U test
RQ4: When encouraged to use ChatGPT for lesson and learning resource generation, what proportion of lessons do teachers use ChatGPT to help with preparation, and does this proportion change significantly as teachers become more familiar with ChatGPT?	Compliance	Teacher weekly diary (weeks 1–10)	Binomial multilevel model
RQ5: When supplied with with the teacher guide on using ChatGPT for lesson and learning resource generation, in how many weeks do teachers consult the teacher guide at least once a) during the first five weeks? b) during the second five weeks?	Implementation fidelity	Teacher weekly diary (weeks 1–10)	Descriptive analysis

Research questions	Methodology area	Data collection methods & participants	Data analysis methods
RQ6 To what extent do teachers adhere to their allocated approaches?	IPE: Adherence	Teach diary, surveys and focus groups; lesson preparation walkthrough	Descriptive statistics (diary, quiz, surveys) Multilevel modelling (moderator analysis) Factor analysis (confidence measures) Thematic analysis (case studies) Content analysis (ChatGPT transcripts)
RQ7 How is ChatGPT used by science teachers while preparing for lessons?	IPE: Fidelity of implementation, adaptation	Teach diary, surveys, quiz and focus groups; lesson preparation walkthrough and ChatGPT transcript	
RQ8 How do teachers use the teacher guide?	IPE: Teacher guide	Teach surveys, quiz and focus groups; lesson preparation walkthrough and ChatGPT transcript	
RQ9 What is the perceived impact of using ChatGPT in lesson and resource preparation?	IPE: Perceived impact, unintended consequences	Teacher surveys and focus groups, senior leader interviews, lesson preparation walkthrough and pupil focus groups	
RQ10 To what extent do moderators affect behaviour and workload changes?	IPE: Moderators	Teach diary, surveys and focus groups; lesson preparation walkthrough	
RQ11 What is usual practice in science teachers' lesson and resource preparation?	IPE: Usual practice	Teach surveys and focus groups; senior leader interviews	

Appendix B



Teacher Guidance: ChatGPT in lesson preparation A Teacher Choices Trial

Thank you for taking part in this research project. This Teacher Choices trial aims to help understand the impact of the use of Generative Artificial Intelligence (GenAI) (specifically ChatGPT) on teaching practice and teacher workload. Half of the schools have been randomly allocated to use ChatGPT and half have been randomly allocated to avoid using any GenAI tools when preparing for lessons.

The research is co-funded by the Education Endowment Foundation (EEF) and the Hg Foundation and independently evaluated by National Foundation for Educational Research (NFER).

This document provides the information you need to take part in the trial. It sets out the details of the approach you have been allocated, with some "do's and don'ts".

For any questions not covered, please email the research team at: GenAI@nfer.ac.uk

You have been allocated to the ChatGPT group

Please read the following guidance carefully.

To ensure the protection of staff, pupils, and their data when using AI, first please read:

- the DfE guidance on [Protecting Data, Pupils and Staff](#) when using GenAI
- the 'What should I watch out for?' section in [Teaching with ChatGPT – Getting Started](#)

Do not enter any pupil or staff personal data into ChatGPT, and please follow your relevant school policies on data protection and/or AI.

Further information is available at the OpenAI Help Centre [Educator FAQ](#)

- ✓ Please follow the guidelines below for the science lessons with the Year 7 and/or Year 8 classes that you nominated for this trial.
- ✓ Please use ChatGPT to support your lesson and resource preparation for 10 weeks, from April 22nd to July 5th.
- ✓ Please read the [Teaching with ChatGPT](https://teachingwithchatgpt.org.uk/) teacher guide (<https://teachingwithchatgpt.org.uk/>) prior to lesson and resource preparation. The teacher guide describes how ChatGPT can support your lesson and resource preparation, and how to get the best results.
- ✓ We recommend you use the first five weeks of the trial to look at, and try out, different parts of the guide. We encourage you to use ChatGPT in your resource preparation as much as you

Restricted

1

can during this time. You can continue to refer to the guide as much or as little as you would like during the second block of five weeks (after May half term).

- ✖ Please don't use any other GenAI tools for your lesson and resource preparation.
- ✓ You don't need to create any additional lesson resources or do any additional planning specifically for this project, over and above what you usually would.

Supporting the evaluation

We have summarised the activities that we will be asking you to complete as part of the research. All are important, but the most vital is completing all weeks of the diary.



Diary: Each week, you will be asked to complete a brief diary about your lesson and resource preparation, including how much time you spent (at any point, including in previous weeks) preparing the Y7/8 science lessons you taught that week. This is really important data for our research and we need all teachers to complete the diary every week. We recommend you keep a record of time spent to help you complete the weekly diary. We also recommend you reply to these questions at the end of each week to support accurate recall. For each participating teacher that completes all ten weeks of the diary, your school will receive £100 to spend at their discretion.



Quiz: Before the May half-term, we will ask you to complete a brief, multiple choice quiz (8 questions) to check your understanding of using ChatGPT in lesson and resource preparation. This quiz will be based on the [Teaching with ChatGPT](#) guidance.



In-person case study visit: If your school is selected, an NFER researcher will visit your school in June to conduct a case study visit. This will involve staff interviews and focus groups, pupil focus groups, as well as a lesson preparation walk-through. If invited, please consider taking part in this activity as your participation and insights are invaluable to our research. Case study schools will receive a £150 thank you payment to spend at their discretion.



Lesson resources: In June, we will ask you to share a copy of the lesson resources for three science lessons that were generated using ChatGPT. These should include a brief lesson outline that details the learning objective(s) alongside any handouts, PowerPoint presentations, pre-prepared interactive whiteboard pages, quizzes, assessments or other text-based resources used in the lesson. You don't need to share your best or worst – a typical example is what we are looking for. These will be reviewed anonymously (we won't attach your name or school to it). We would be grateful if you could share these resources when asked.

For each participating teacher who shares their lesson resources, your school will receive £30 to spend at their discretion.



ChatGPT transcript: In June, we will also ask you to share any ChatGPT transcripts that were used in the preparation of these lesson resources. We would be grateful if you could share these resources when asked.



End-of-trial online teacher survey: In June–July, we will ask you to complete an online teacher survey to share your views and confidence in relation to your science teaching. Please complete this survey when sent the link.

Timeline



Below is an outline of planned research activities over the next few months. We'll contact you for each of these research activities, which are essential for our research progress. We greatly appreciate your ongoing participation in the research, and we look forward to your continued support.

Date	Research activities	Mode of completion
Every Thursday starting from 25 th April for 10 weeks (except May half-term)	Complete your online weekly diary (3 minutes) by the end of that week	Online via email link
16 th May – 4 th June	Complete the short online quiz (5 minutes)	Online via email link
June	In-person research activities (schools selected for the case study sample)	NFER researcher's visit to your school
June	Send NFER a copy of the lesson resources for three science lessons generated using ChatGPT Send NFER your ChatGPT transcripts	Upload material on NFER secure portal or send these via email
End-June – Mid-July	Complete the end-of-trial online teacher survey (Up to 15 minutes)	Online via email link

Thank you for your participation in our study; your contribution is immensely valued.

Teacher Guidance: ChatGPT in lesson preparation A Teacher Choices Trial

Thank you for taking part in this research project. This Teacher Choices trial aims to help understand the impact of the use of Generative Artificial Intelligence (GenAI) (specifically ChatGPT) on teaching practice and teacher workload. Half of the schools have been randomly allocated to use ChatGPT and half have been randomly allocated to avoid using any GenAI tools when preparing for lessons.

The research is co-funded by the Education Endowment Foundation (EEF) and the Hg Foundation and independently evaluated by National Foundation for Educational Research (NFER).

This document provides the information you need to take part in the trial. It sets out the details of the approach with some "do's and don'ts".

For any questions not covered, please email the research team at: GenAI@nfer.ac.uk

You have been allocated to the **Non-GenAI** group

Please read the following guidance carefully.

- ✓ Please follow the guidelines below for the science lessons with the Year 7 and/or Year 8 classes that you nominated for this trial.
- ✗ Do not use any GenAI tools (such as ChatGPT, Gemini, or teaching specific AI tools) for your lesson and resource preparation for 10 weeks from April 22nd to July 5th.
- ✓ When preparing for lessons and creating resources to use, please feel free to draw on the sources you already use e.g. other teachers, departmental shared resources, textbooks, external schemes of work, teacher websites/forums.
- ✓ You don't need to create any additional lesson resources or do any additional planning specifically for this project, over and above what you usually would.

We realise that you may have signed up to the project because you are already using, or are interested to try using, GenAI/ChatGPT. Having a group that is not using GenAI for their lesson preparation is really important for understanding the impact on workload and we appreciate you following this approach for the trial period.

Supporting the evaluation

We have summarised the activities that we will be asking you to complete as part of the research. All are important, but the most vital is completing all weeks of the diary.



Diary: Each week, you will be asked to complete a brief diary about your lesson and resource preparation, including how much time you spent (at any point, including in previous weeks) preparing the Y7/8 science lessons you taught that week. This is really important data for our research and we need all teachers to complete the diary every week. We recommend you keep a record of time spent to help you complete the weekly diary. We also recommend you reply to these questions at the end of each week to support accurate recall. For each participating teacher that completes all ten weeks of the diary, your school will receive £100 to spend at their discretion.



In-person case study visit: If your school is selected, an NFER researcher will visit your school in June to conduct a case study visit. This will involve staff interviews and focus groups, pupil focus groups, as well as a lesson preparation walk-through. If invited, please consider taking part in this activity as your participation and insights are invaluable to our research. Case study schools will receive a £150 thank you payment to spend at their discretion.



Lesson resources: In June, we will ask you to share a copy of the lesson resources for three science lessons. These should include a brief lesson outline that details the learning objective(s) alongside any handouts, PowerPoint presentations, pre-prepared interactive whiteboard pages, quizzes, assessments or other text-based resources used in the lesson. You don't need to share your best or worst – a typical example is what we are looking for. These will be reviewed anonymously (we won't attach your name or school to it). We would be grateful if you could share these resources when asked. For each participating teacher who shares their lesson resources, your school will receive £30 to spend at their discretion.



End-of-trial online teacher survey: In June–July, we will ask you to complete an online teacher survey to share your views and confidence in relation to your science teaching. Please complete this survey when sent the link.

Timeline




Below is an outline of planned research activities over the next few months. We'll contact you for each of these research activities, which are essential for our research progress. We greatly appreciate your ongoing participation in the research, and we look forward to your continued support.

Date	Research activities	Mode of completion
Every Thursday starting from 25 th April for 10 weeks (except May half-term)	Complete your online weekly diary (3 minutes) by the end of that week.	Online via email link
June	In-person research activities (schools selected for the case study sample)	NFER researcher's visit to your school
June	Send NFER a copy of the lesson resources for three science lessons	Upload material on NFER secure portal or send these via email
End-June – Mid-July	Complete the end-of-trial online teacher survey (Up to 15 minutes)	Online via email link

Thank you for your participation in our study; your contribution is immensely valued.

Appendix C

INTRODUCTION – Show to all
<div><p>NFER National Foundation for Educational Research</p></div>
<p>Teacher Diary – ChatGPT in lesson preparation</p> <p>[show allocated approach name:] You are in the: [ChatGPT /non-GenAI] group</p> <p>As part of this research project, we are asking you to complete a weekly diary for 10 weeks during the summer term. We are very grateful for your support.</p> <p>You will be asked the same questions each time, relating to each week in turn. You will need to complete the diary entry for any previous incomplete weeks before you can move on to the next week's questions. The diary will become available to you each Thursday – we will email you to let you know it is ready for that week's entries. Please answer all of the questions every week.</p> <p>It should take no more than 3 minutes to respond to these diary questions every week.</p> <p>If you have any queries about this diary, please contact GenAI@nfer.ac.uk.</p> <p>The information you provide in this diary will only be used as part of the research and nobody at your school will see your answers. We will not identify individual teachers or schools in the report. You can find more details about the research and how we will use the data you provide on the project information site.</p> <p>Please use the buttons at the bottom of the page to move through the diary, please <u>do not</u> use your browser's forward and back buttons.</p> <p>Please note that if the diary is left inactive for over 20 minutes you will be timed out and will need to click on the link again. If you only partially complete the diary (including full or partial weeks), any answers that you have given will still be analysed.</p> <p>Once you submit responses for each week, you will not be able to go back and change any of your answers for that week.</p>

At the end of Spring term, you completed a teacher survey where you nominated XX Year 7 science classes and XX Year 8 science classes for this research project.

Please answer the following questions in relation to these classes.

Q1 –Open [numeric] Response, whole numbers only, Ask all, Force [for the pop up message please say: this question is important for us to understand the impact on teacher workload. Please provide your best estimate]

How many **science lessons** did you teach for your nominated science classes during the week commencing 22nd April?

If you were not working or did not teach any science lessons for any reason this week, please enter 0 (zero)

(numeric response) ____ lessons

If Q1 = 0 (zero) please show this screen/message:

You have reported that you did not teach any science lessons in the week of 22nd April.

If this is correct, please click submit to close this week's diary.

If this is not correct and you did teach some science lessons in the week of 22nd April, please use the back button below to go back and change your answer.

Show "Back" button and "Submit" button

[If they click submit then route to the end of this week's diary and then send the following week as normal]

Q2 –Open [numeric] Response, limit to 1 decimal place, Ask all, Force [for the pop up message please say: this question is important for us to understand the impact on teacher workload. Please provide your best estimate]

Approximately how many hours did you spend on lesson or resource preparation for the [\[prepopulate with number from Q1\]](#) lessons that you taught your nominated classes during the week commencing 22nd April?

Include tasks that took place during weekends, evenings or other out of class hours.

Exclude all time spent teaching.

Do not include time marking or doing administrative tasks.

An estimate is sufficient. If you did no lesson or resource preparation for the lessons you taught in the week commencing 22nd April, enter 0 (zero). Round to the nearest half hour. As an example, three and a half hours would be recorded as 3.5 below.

(numeric response) ____ hours

<p>Q3a –Open [numeric] Response, whole numbers only (validation: Q3a<=Q1), Ask ChatGPT arm only.</p> <p>Force [for the pop up message please say: this question is important for us to understand the impact on teacher workload. Please provide your best estimate]</p>	
<p>Please continue to think about the [prepopulate with number from Q1] lessons taught during the week commencing 22nd April. For how many of these lessons did you use ChatGPT for lesson and resource preparation?</p> <p><i>An estimate is sufficient. Include all lessons for which you accessed ChatGPT during your preparation even if it was only for a small part of the preparation. If you did not use ChatGPT when preparing this week's lessons, please enter 0 (zero).</i></p>	
<p>(numeric response) ____ lessons</p>	

<p>Q3b–Open [numeric] Response, whole numbers only, (validation: Q3b<=Q1), Ask Non GenAI arm only.</p> <p>Force [for the pop up message please say: this question is important for us to understand the impact on teacher workload. Please provide your best estimate]</p>	
<p>Please continue to think about the [prepopulate with number from Q1] lessons taught during the week commencing 22nd April. For how many of these lessons did you use any GenAI tools for lesson and resource preparation?</p> <p><i>Although you are in the Non GenAI group, we still need to ask you whether you used any GenAI tools, just in case! Please answer honestly, thank you.</i></p> <p><i>An estimate is sufficient. Include all lessons for which you accessed GenAI tools during your preparation even if it was only for a small part of the preparation. If you did not use any GenAI tools when preparing this week's lessons, please enter 0 (zero).</i></p>	
<p>(numeric response) ____ lessons</p>	

Q4 – Single response, Ask ChatGPT arm only, nudge

4.	Did you use the 'teachingwithchatgpt.org.uk' guidance during the week commencing 22 nd April?	Please select one	3.1	Yes
			3.2	No

CLOSING STATEMENT – Show to all**Weekly diary entry complete**

Thank you for completing this weekly diary. The new diary questions will appear on next Thursday.

Thank you for all your support with this research project.

Weekly diary entry complete

Thank you for completing the weekly diary. You have now completed all diary entries for this research project.

We hugely appreciate all your support.

Appendix D

Randomisation code. Run in R version 4.2.2 on 19/03/2024.

```
## EEAI randomisation - school level

# set working directory and load libraries
rm(list=ls())
setwd(choose.dir())

for(x in c("readxl", "ggplot2")){
  if(!x%in%rownames(installed.packages())){install.packages(x)}
  library(package=x, character.only=T)
}
rm(x)

# load randomisation file - downloaded 19/03/2024 from (sharepoint
link)
df<-data.frame(read_excel("EEAI_Data for Randomisation.xlsx"))
str(df)

ggplot(df, aes(x=Number.of..Teachers.completed.baseline.survey..per.s
chool.))+

geom_histogram(binwidth=1, center=1, fill="grey50", colour="black")+
  scale_x_continuous(name="Number of Teachers completed baseline
survey (per
school)", breaks=seq(0, 2*ceiling(max(df$Number.of..Teachers.completed
.baseline.survey..per.school.)/2), by=2))+
  theme_bw()
table(df$Number.of..Teachers.completed.baseline.survey..per.school.)
# decision by team to stratify by number of teachers
# two bins - <=6 and >=7

df$NumberOfTeachers_Binned<-factor(c("1-6", "7-
10")[1*(df$Number.of..Teachers.completed.baseline.survey..per.school
.>=7)+1])
table(df$NumberOfTeachers_Binned, df$Number.of..Teachers.completed.ba
seline.survey..per.school.)

# randomise
set.seed(as.numeric(as.Date("2024-03-19")))
df$RandomisationGroup<-NA
for(i in levels(df$NumberOfTeachers_Binned)){
  df[df$NumberOfTeachers_Binned%in%i,]$RandomisationGroup<-
sample(rep(LETTERS[1:2], each=ceiling(sum(df$NumberOfTeachers_Binned%
in%i)/2)))[1:sum(df$NumberOfTeachers_Binned%in%i)]
}
rm(i)

# check balance
table(df$RandomisationGroup)
```

```

table(df$RandomisationGroup,df$NumberOfTeachers_Binned)
table(df$RandomisationGroup,df$Number.of..Teachers.completed.baselin
e.survey..per.school.)
sum(df[df$RandomisationGroup%in%"A",]$Number.of..Teachers.completed.
baseline.survey..per.school.)
sum(df[df$RandomisationGroup%in%"B",]$Number.of..Teachers.completed.
baseline.survey..per.school.)
table(df$RandomisationGroup,df$EstablishmentTypeGroupName)
summary(df[df$RandomisationGroup%in%"A",]$NumberOfPupils)
summary(df[df$RandomisationGroup%in%"B",]$NumberOfPupils)

# assign groups
df$RandomisationGroup<-sample(c("ChatGPT","Non-
GenAI"))[factor(df$RandomisationGroup)]

# write randomisation group to csv
write.csv(df,"RandomisationOutput.csv")

```

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>



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

<https://educationendowmentfoundation.org.uk>

 @EducEndowFoundn

 Facebook.com/EducEndowFoundn