Department for Education

Approaches to developing the education technology (EdTech) impact testbed

Scoping report

December 2025

Contents

Introduction

The UK government is committed to leveraging technology to support every child and young person to achieve and thrive.

Technology, including Artificial Intelligence (AI), offers a powerful way to drive positive change, improve outcomes, and create efficiencies and benefits across the education system.

In England, schools and colleges have autonomy over their EdTech purchasing decisions. To choose the right EdTech for their setting, they need to be able to navigate the EdTech market with confidence.

The Department for Education (DfE) has been working to support informed decision-making, introducing digital and technology standards that promote safe, cost-efficient practices and open new learning opportunities for students. In 2025, the DfE also launched Plan Technology for Your School, a digital service designed to help schools make strategic decisions about what technology to buy and how to implement it effectively.

However, schools and colleges want to know what works, through access to evidence that clearly demonstrates if EdTech solutions deliver measurable impact and value for money. Recent surveys reinforce this need for evidence. In a Teacher Tapp poll, 87% of teachers responded that they wanted clear proof of EdTech Effectiveness. However, there is a lack of evidence available to educators, recent sector consultation and research into early adopters of AI highlighted evidence gaps on the impact of EdTech and AI on learning outcomes and workload reduction.

The DfE's Technology in Schools Survey 2024 to 2025 reports that school budget remains the most influential factor in technology investment decisions, cited by 91% of school leaders, followed by evidence of best practice (76%). Given the tight fiscal context schools and colleges are currently operating, it is important to ensure that the EdTech adopted by schools and colleges is safe, supports teachers and learners, and has tangible benefits – something that is not guaranteed or known due to lack of evidence.

Decision-makers in schools and colleges need to be able to make informed decisions on the technology they purchase and deploy that align with the priorities and goals of their school or college. Quality evidence needs to be available and accessible to support informed decision making. However, industry has struggled to produce accessible high-quality evidence on the efficacy and impact of EdTech products and practices.

Schools and colleges face pressures to respond and consider adopting new technologies, and without timely evidence, they risk investing in tools that may not deliver promised benefits or could even introduce unintended harms. The complexity of real-

world education settings makes generating quality evidence challenging. EdTech products, especially AI tools, are evolving faster than traditional research cycles can keep up with, so there is a need to understand more about how to generate high quality evidence in this transient EdTech landscape.

The impact of an EdTech tool is rarely singular or independent, there are usually a multitude of impacts and factors, including enabling infrastructure and user capability, that need to be considered and in place for technology to deliver genuinely positive and equitable benefits for educators and learners.

Closing the gap between EdTech developers and educators is also key to creating solutions that are grounded in evidence, aligned with pedagogy, and responsive to real classroom needs.

The DfE is uniquely positioned to bring together stakeholders across the education sector, shape market and sector practice, and drive and accelerate an evidence-based approach to EdTech. To help address these challenges and close these gaps, the DfE are running an EdTech Evidence Board pilot exploring the best way to set quality standards for current and future products and support high quality evidence production. The pilot is trialling evidence assessment with a small number of products this year to establish how to apply evidence criteria to understand what products are adding value for teaching and learning.

The DfE has also committed to funding an EdTech Impact Testbed. This will provide a structured environment for testing and evaluating EdTech products and practices, including artificial intelligence and assistive technologies, aiming to drive a more evidence-based approach to edtech development and purchasing.

Report outline

This report was commissioned by the DfE to explore potential approaches to developing an EdTech testbed, with the aim to test EdTech in schools and colleges to generate evidence of impact.

The report presents an overview of existing EdTech testbeds in education settings and a range of possible design approaches to meet set objectives and priorities. It also outlines the benefits and trade-offs of selecting different education outcomes and school-based tasks to test. The report also explores how the pilot could measure impact and what evaluation considerations the DfE should consider when deciding on the pilot's design.

The research was conducted by the Open Innovation Team (OIT).

What is an EdTech testbed?

EdTech refers to the use of digital tools and resources to improve teaching, learning and school administration and management. It encompasses a wide range of technologies, including hardware, software, services and AI-based tools.

A testbed is a structured environment where new EdTech can be tested in real-world educational settings. It brings together education settings, researchers, and developers to create a framework to host individual studies to solve a particular problem.

Individual studies are conducted within the testbed environment. Each study can have different research questions, methodologies and outcomes. Multiple studies can run simultaneously, testing different tools. The testbed infrastructure typically outlasts the individual studies it hosts.

Research objectives and priorities

The Open Innovation Team were asked to deliver a series of objectives through this project:

- To conduct research into EdTech testbed programme models, using existing examples from the UK and internationally, to determine the current research gaps and identify the characteristics of successful EdTech testbeds to develop the evidence base and deliver benefits for stakeholders.
- To present a realistic programme design and methodology for the DfE's EdTech Impact Testbed, including testbed delivery and evaluation.

The DfE outlined three main objectives for the EdTech Impact Testbed:

- **Identifying EdTech tools.** The pilot testbed and first study will be used to identify promising technologies that have the potential to enhance teaching and learning.
- **Generating evidence.** The testbed will be used to support EdTech organisations and education settings in collecting evidence on the impact of their tools and use of technology. The evidence will be used to help schools and colleges make informed decisions about appropriate tools to address their needs.
- **Testing the testbed.** A pilot phase of the testbed and first studies will be used to explore how best a testbed can be designed to deliver on each of these aims.

The DfE have identified three priority areas to test EdTech tools aligning with departmental mission delivery and supporting all children and young people to achieve and thrive.

The areas are:

- **Teacher and administrative workload.** The DfE is committed to removing unnecessary burdens and supporting schools and colleges to introduce more efficient practices, so that staff can focus on activities that most directly improve learner outcomes.
- **Learning outcomes.** A key aim of advancing EdTech tools is to help raise attainment and progress for all learners, particularly those from disadvantaged backgrounds.
- **Inclusion of all children, including those with special educational needs and disabilities (SEND).** EdTech has the potential to enhance mainstream educational outcomes and experiences for learners, including those with SEND by improving accessibility and support.

Research approach

To design options for testing EdTech products and practice in real education settings to generate evidence of impact, the Open Innovation Team brought together evidence from a range of sources:

- **Expert interviews.** Interviews with 12 experts from academia and the EdTech sector. The majority of participants were based in the UK, as well as experts from Finland, the Netherlands and the United States.
- **Facilitated two workshops.** Two workshops with experts to focus on some of the specifics of the literature and to test early design options.
- **Conducted a literature review.** An initial review of academic and grey literature on testbeds at the project's outset, followed by a second review midway to inform our focus.
- **Launched an Expression of Interest.** To build a pipeline of schools, colleges and EdTech companies to participate in the pilot study, the DfE also commissioned OIT to launch two Expressions of Interest (EoIs). The EoIs invited schools and colleges, and separately EdTech suppliers, to register their interest in the pilot and provide key information such as type of EdTech currently used in their school, and for suppliers, what solutions they provide. These opened on 10th July 2025 and closed 1st September 2025.

EdTech testbed models

This section outlines what an EdTech testbed is and what it can achieve. It then examines the four main approaches to structuring an EdTech testbed, analysing their strengths and limitations.

EdTech testbeds create an infrastructure to host individual studies

An EdTech testbed is a structured environment where new education technology (EdTech) can be tested in real-world educational settings[1]. An EdTech testbed acts as a platform or ecosystem to bring together schools, colleges, researchers, and developers. The goal of a testbed is to generate evidence about what EdTech tools work, for whom, and under what conditions. It helps address the evidence gap in EdTech by enabling systematic and independent testing of products.

Testbeds form the infrastructure that can host individual studies, which are conducted within the testbed environment. These studies can have different research questions, methodologies and outcomes. Multiple studies can run simultaneously, each testing different tools, hypotheses, or user groups, and the testbed will generally outlast the studies that it hosts.

Examples of EdTech testbeds

The examples featured are two large EdTech testbeds in the Netherlands and USA. Both testbeds are well-established, have received substantial funding over multiple years, and employ full-time staff ranging from 12 to 40 members, and work with learners up to eighteen. The scope of these examples exceeds that of the DfE's EdTech Impact Testbed Pilot. They have been selected to illustrate how testbeds can take different approaches to testing.

[1] Towards Systemic EdTech Testbeds: A Global Perspective.

The National Education Lab AI, Netherlands

The National Education Lab AI (NOLAI), based at Radboud University in the Netherlands, is a national innovation hub for responsible and effective use of AI in education. It is a 10-year project funded by the National Growth Fund and the European Union's NextGenerationEU initiative. It has a budget of €80 million from 2022 to 2032, and an additional €63 million to support the scale-up of prototypes.

NOLAI focuses on creating prototypes and researching their pedagogical, technical, and ethical impacts using a living-lab model. It has two main programmes:

- Co-creation programme. Launches AI prototypes, working with teachers, researchers and EdTech companies. Rapid evaluation cycles or design-based research are conducted, followed by classroom testing and iterative refinement.
- Scientific programme. Deploys AI tools directly in schools, and studies them on pedagogy, teacher training, ethics, AI tech, and data sustainability.

Each project starts by identifying a need, and successful prototypes enter the implementation stage for scale-up. There are currently 17 active projects within this testbed.

Leanlab Education, USA

Leanlab Education is a non-profit based in Kansas City which guides EdTech tools from ideation to classroom integration. Their mission is to deploy EdTech solutions that are co-designed with schools to leverage collaborative, research-backed development. The research framework is divided into 4 tiers:

- Usability Studies. These are typically 3-4 weeks where participants use tools and researchers observe and interview them to assess usability.
- Feasibility Studies. These are typically 4-8 weeks where tools are deployed in classroom settings. Integration and impact are assessed using journals and focus groups.
- Rapid-Cycle Evaluation. These typically last around 8-12 weeks over 2-3 cycles and involve iterative testing and refinement where teachers implement tools, give feedback, and developers make updates between cycles.
- Codesign Product Research/School Network Pilots. Teachers pilot tools in classrooms, where schools can get grants.

The testbed has two main programmes: the Agile Network (involving over 140 schools and 80,000 students), and the Codesign Collective (involving over 475 educators from 37 states).

What objectives can a EdTech testbed achieve?

EdTech testbeds generally have multiple objectives. These include:

- **Generating evidence on EdTech tools through individual studies.** Depending on how they are designed, testbeds can produce different kinds of evidence.
- **Supporting the EdTech sector to develop more effective tools.** Developers receive structured feedback from educators and/or learners, enabling iterative improvements and alignment with classroom needs.
- **Fostering collaboration between schools and colleges, researchers, and EdTech companies.** Testbeds can act as convening spaces for co-design and shared learning.
- **Reducing risk in EdTech adoption.** Education institutions can trial tools in a low-stakes environment before committing to full-scale implementation.
- **Identifying tools that are ready to be scaled across schools and colleges.** By testing tools in diverse contexts, testbeds help determine which products are ready for wider adoption and which need further development.
- **Providing guidance to educators on how to choose and adopt effective EdTech tools.** Testbeds can translate evaluation findings into practical insights, helping schools and colleges to make informed procurement and implementation decisions.
- **Supporting the professional development of educators.** Participation in testbeds often includes training and reflective practice, enhancing educators' digital confidence and pedagogical skills.
- **Building a culture of evidence use in education.** By embedding evaluation into everyday practice, testbeds help normalise the use of data and research in decision-making.
- **Supporting equity and inclusion in EdTech access.** Testbeds can be designed to include a diverse range of education providers and learners, ensuring tools are tested against a range of needs.

What are the different EdTech testbed design models?

An EdTech testbed can take a variety of forms, depending on the objectives it is seeking to achieve. The first step in designing an EdTech testbed is to choose the appropriate testbed model. By choosing a model, the testbed's individual study designs can be tailored to meet specific objectives. In 2019, Nesta conducted a literature and case study review of different testbed approaches. They identified four different ways of structuring

an EdTech testbed. It should be noted that these models are not mutually exclusive, and an EdTech testbed can draw on components from each simultaneously.

Co-design

This model enables teachers and learners to collaborate with EdTech companies to identify needs, develop prototypes and design new studies for measuring impact. It creates space for all stakeholders to explore how innovative technology can enhance learning outcomes and experiences. Common methods include design ethnography, contextual design, and storyboarding. An example of a co-design model is MindCET.

Key benefits of this model include:

- **They are best suited for tools in early development.** This includes bringing developers and potential users together to develop new ideas, tools and prototypes, alongside designing studies for testing them.
- **They empower teachers and learners to shape products.** Encourages user-centred design that aligns with classroom needs, incentivising educators and learners to engage.
- **They enable long-term developer–educator partnerships.** As educators and developers work closely together on the design of new EdTech, they often build relationships that outlast the initial project.

Key challenges include:

- **They are limited in generating generalisable impact evidence.** As these projects are typically small scale, directly involving a small number of collaborators, they are often context-specific and may not address wider issues.
- **Time- and resource-intensive.** This approach often requires specialist facilitation and extended engagement to produce the best results.

Test and Learn

This model aims to take a more structured, consistent approach to studies within the testbed programme. This focuses on rapid, real-world trials of EdTech tools in education settings to assess their effectiveness. It typically uses mixed-methods evaluation approaches (quantitative and qualitative) and supports iterative improvements based on findings. The goal is to generate actionable insights quickly and affordably. An example of a test and learn model is iZone, a project from New York City Department of Education's Short-Cycle Evaluation Challenge. In 2016, the Centre for Children and Technology, New York published a report on iZone's test and learn model. The report found the model proved effective for rapid technology evaluation but revealed the critical

importance of context, implementation quality, and multi-stakeholder feedback in determining the success of a tool.

Key benefits of this model include:

- **Best suited for tools in mid-stage development.** Enables testing of tools that are ready for classroom testing.
- **Generates practical, actionable insights**. Helps inform product development and decision making in real-world education settings. These early indicators can help schools and colleges decide whether to adopt a product.
- **Produces some evidence on 'the promise' of impact.** However, this may not meet the standards required for high-level impact evaluations.
- **Provides feedback for EdTech companies to improve their tools.** Feedback from trials can be used to refine and enhance tools.

Key challenges include:

- **Limited depth of evidence.** While this model prioritises speed and practicality, it may not provide more rigorous evidence of impact.
- **Resource demands for coordination and analysis.** Even with smaller-scale testing, centrally managing multiple trials (especially with different tools) can require significant capacity.

Evidence Hub

This approach centres on generating, synthesising, and sharing evidence about the performance of EdTech tools. By providing data on product impact, it helps educators make informed decisions and allows developers to demonstrate the effectiveness of their tools. Common methods include quasi-experimental impact studies, randomised controlled trials and user research. An example of an evidence hub model is the Education Endowment Foundation (EEF).

Key benefits of this model include:

- **Generates high-quality, independent evidence of impact.** Supports evidence-based decision making for educators, policymakers and developers.
- **Reduces duplication and guides future research.** Helps prioritise which tools to test and avoids repeated evaluation of the same products.
- **Useful for schools and colleges with limited evaluation capacity.** Insights from studies in this model, if shared freely, can support procurement and policy decisions when in-house evaluation isn't feasible.

Key challenges include:

- **May overlook local and contextual factors**. Findings might not fully reflect the specific needs of individual schools or settings.
- **Requires robust and often costly methods**. High-quality data collection and analysis can be complex and resource intensive.

EdTech Network

Unlike the other models that emphasise developing or testing individual EdTech products, this approach focuses on sharing insight and can be applied across all stages of EdTech product development. This model aims to build networks of educators, developers and policymakers that can share their experiences and best practices. It promotes peer learning, collaborative evaluation and collective problem-solving. Common methods include webinars, working groups, and conferences. An example of an EdTech network model is Digital Promise League of Innovative Schools.

Key opportunities of this model include:

- **Encourages professional development and builds the skills, confidence and knowledge of educators.** By participating in a networked learning community, educators are exposed to new tools, pedagogies, and use cases, helping them become more informed and empowered users of EdTech.
- **Promotes knowledge sharing between schools and colleges.** Helps adapt and scale effective practices across institutions.
- **Best for scaling practice and building capacity.** Ideal for spreading innovation, sharing promising practices and identifying the most promising EdTech products.

Key challenges include:

- **Needs sustained coordination.** Long-term success depends on active engagement and ongoing communication.
- **Impact may vary across participants.** Not all members may benefit equally.

An overview of some of the typical kinds of evidence each testbed model produces, with some examples, can be found in Annex 6.

Comparing testbed models against objectives

This section analyses how the different testbed models align with the three priority objectives, and which approach may best fulfil these.

The three priority objectives for the testbed initiative are:

- **Identifying EdTech tools.** The pilot testbed will be used to identify promising technologies that have the potential to enhance teaching and learning.
- **Generating evidence.** The testbed will be used to support EdTech organisations and education settings in collecting evidence on the impact of their tools and use of technology. The evidence will be used to help schools and colleges make informed decisions about appropriate tools to address their needs.
- **Testing the Testbed.** The pilot testbed will be used to explore how best a testbed can be designed to deliver on each of these aims.

Table 1 provides an overview of the degree to which each of the different testbed models can satisfy these outcomes. It also considers other objectives that can be achieved with a testbed.

Table 1: A summary of how each testbed model structure addresses key objectives and other goals[2]

Objective	Co-design	Test and learn	Evidence Hub	EdTech Network
DfE key objective: Identify EdTech tools	Early-stage – Supports prototype development with educators	Mid-stage – Tests pilot-ready tools for refinement	Mature-stage – Focus on tools ready for evaluation	All – Shares insights from tools used at all stages
DfE key objective: Generate evidence on impact	Limited – High-level feedback during early development	Yes – Structured trials inform evidence base	Yes – Formal evaluation methods embedded	Somewhat – Promotes discussion but lacks formal methods

[2] Table key: white squares = fully meets objective, light grey squares = somewhat meets objective depending on design, dark grey = limited ability to meet objective.

Objective	Co-design	Test and learn	Evidence Hub	EdTech Network
DfE key objective: Test the testbed	Variable – Depends on broader pilot design	Yes – Multiple cycles can be used to explore different approaches	Variable – Depends on broader pilot design	Variable – Depends on broader pilot design
Support educator development	Somewhat – Involves teachers in design processes	Somewhat – Focus is tool testing, with some training	Somewhat – Involvement in studies builds capacity	Yes – Networking and training enhance practice
Aid better products	Yes – Early input shapes design direction	Yes – Feedback loops improve usability	Yes – Evaluation informs product quality	Yes – Insight from users supports iteration
Support the sector	Yes – Direct co-creation with educators	Yes – Rapid testing and feedback accelerates growth	Somewhat – Evidence informs long-term dev	Yes – Peer exchange sustains improvement
Foster collaboration	Yes – Design partnerships between schools and developers	Somewhat – Collaboration occurs during trials	Yes – Central hub aligns stakeholders	Yes – Builds long-term, trust-based relationships
Reduce adoption risk	Somewhat – Reduces mismatch risk, but tools may be immature	Yes – Trials support informed decisions	Yes – Stronger evidence lowers uncertainty	Somewhat – Peer learning helps identify risks
Identify scalable tools	Rarely – Tools not yet market-ready	Somewhat – May surface promising tools that need further testing	Yes – Evaluation supports scale-up	Yes – Networks amplify proven tools

Objective	Co-design	Test and learn	Evidence Hub	EdTech Network
Build evidence culture	Limited – Focus on design, not evidence practices	Somewhat – Introduces evaluation habits	Yes – Embeds systematic evidence use	Somewhat – Encourages reflective dialogue
Support equity and inclusion	Yes – Inclusive design addresses diverse needs	Variable – Depends on diversity of testing contexts	Variable – Inclusion tied to scope of evaluation	Yes – Promotes inclusive access via community norms

While each model brings distinct strengths, no single approach can fully satisfy all of the outlined objectives. Also, it is crucial to note that these approaches are not mutually exclusive. A combination of any, or all, of these models is feasible and, in some cases, may be preferable to a single approach.

Selecting an outcome to test and evaluate

An EdTech testbed can have a broad focus looking at a single, or multiple, objectives simultaneously. However, individual studies within the testbed require a more well-defined focus that limits scope. Narrowing the scope of a study helps ensure rigour and feasibility. It also informs the selection of tools, the allocation of participants, and decisions on what data could be collected and how.

To guide the focus of the EdTech Impact Testbed, the DfE have identified three priority areas for testing and evaluating EdTech tools:

- **Teacher and administrative workload.** The DfE is committed to removing unnecessary burdens and supporting schools and colleges to introduce more efficient practices, so that staff can focus on activities that most directly improve learner outcomes.
- **Learning outcomes.** A key aim of advancing EdTech tools is to help raise attainment and progress for all learners, particularly those from disadvantaged backgrounds.
- **Inclusion of all children, including those with special educational needs and disabilities (SEND).** EdTech has the potential to enhance mainstream educational outcomes and experiences for learners, including those with SEND by improving accessibility and support.

Each of these areas represents a significant policy priority. However, assessing the impact of EdTech tools in any of these areas requires careful consideration of how relevant outcomes can be measured within the constraints of the initial study. Narrowing scope will not only make the tools easier to evaluate but will also facilitate easier delivery and implementation of the chosen EdTech tools.

This section examines how each of the three priority areas could be addressed and the practical challenges involved. Finally, explaining why initial pilot studies could focus on workload reduction.

Practical considerations for addressing priority outcomes

While testing all three outcomes might seem to maximise the value of a pilot by addressing multiple high-priority areas, doing so would introduce major methodological and logistical challenges that could dilute the reliability and clarity of any evidence collected. These challenges include:

- **Timeframe constraints**. A limited 6–12-week timeframe makes it extremely difficult to capture reliable evidence for complex or longer-term outcomes, such as

improved learner attainment or enhanced inclusion of SEND students. These outcomes typically require extended observation, follow-up, and more nuanced data collection methods.

- **Increased burden on participants**. Teachers and learners would face higher demands, such as learning to use multiple tools, completing more extensive monitoring tasks, and delivering additional assessments. This could reduce engagement and compromise data quality.
- **Challenges with tool selection.** Most EdTech products are designed to target specific challenges. Therefore, addressing all three outcomes would require selecting a larger number of tools to participate in the study. This would require more resources (to recruit a larger cohort for testing) and increase the complexity of managing and delivering the evaluation (e.g. through bespoke training for each tool or collecting different measures for each outcome).

Focus of initial studies

Focusing the initial pilots on only one priority outcome: reducing teacher and staff workload, would support managing the scope and effective use of funds. This focus would help in two ways:

- **A more manageable design.** Including tools targeting all three priority outcomes (workload, student learning, inclusivity) would increase complexity, cost, and delivery time. A single-outcome focus allows for a more streamlined and manageable pilot.
- **Short-term indicators of impact.** Given the limited testing window (6-12 weeks), workload-related impacts are more readily observable than changes in student learning or inclusivity outcomes, which will require longer timeframes to accurately assess.

Focus of future studies

Focusing the initial studies on workload reduction not only aligns with a critical policy objective but also increases the likelihood of generating actionable insights that can inform the DfE's broader EdTech strategy. Importantly, this approach does not preclude the exploration of other outcomes in future phases of the testbed programme, where longer timeframes and larger sample sizes could support robust evaluations of EdTech's impact on learning or inclusion. Expanding the focus would also help:

- **Fill current evidence gaps**. Particularly in relation to the impact of EdTech on students with SEND, where experts noted there is a current lack of evidence.

- **Address priority areas for schools and colleges.** Experts noted that both improving learner outcomes and inclusivity of students with SEND are currently, and will likely remain, key priority areas for all educational institutions.
- **Facilitate more suitable study designs.** While it may not be feasible to capture the full impact of EdTech tools on complex outcomes (like inclusion and learning outcomes) within a 6–12-week timeframe, future studies could be designed to allow extended follow-up, enabling a better understanding of sustained impacts.

Selecting a task to test

If initial pilot studies focus on reducing teacher and administrative workload, the next decision requires selecting which specific teaching and/or administrative task to test. This will enable the study to assess EdTech tools that address the most pressing workload challenges educators face.

This section reviews three main categories of workload reduction tools (lesson planning, marking and feedback, and administrative tasks) to determine which offers the greatest potential for meaningful impact within the initial study's constraints. The analysis reviews the existing evidence base on the relevance of these tools to current teacher pressures, their potential time savings and presence of other studies.

Categorising teacher and administrative tasks

Following a review of the EdTech market, it's possible to conceptualise the tools that address teacher workload across three distinct categories: lesson planning, marking and feedback, and administration. These categories reflect where teachers spend their non-teaching time and potentially represent the most significant opportunities for workload reduction.

On average across OECD education systems, about half of teachers surveyed report excessive administrative work as a source of work-related stress. The UCU 2021 workload survey, reports increased administrative work remains the biggest contribution to change in workload for staff in further education colleges. Widening of duties, increased student numbers and the reduction of staff also remain significant contributors to change in workload.

DfE's own data, Working lives of teachers and leaders report, demonstrates the overall time burden placed on teachers, driven largely by non-teaching tasks. In 2025, full-time teachers worked an average of 50.1 hours per week, whilst the average time spent teaching for those with teaching responsibilities was 23.3 hours per week. Full-time leaders worked, an average of 56.5 hours, and only just over one-in-four (26%) teachers and leaders agreed that they had an acceptable workload in 2025. General administrative work[3] has consistently been the area that the largest proportion of teachers and middle leaders say they spend too much time on. This data highlights the need for effective interventions to reduce the time teachers spend on these essential but time-intensive tasks.

[3] General administrative work includes tasks such as communication, paperwork, work emails, and other clerical duties they undertake in their job as a teacher.

Lesson planning

EdTech tools designed to support lesson planning have evolved significantly in recent years, shifting from basic resource hubs to intelligent planning assistants.

Key considerations for lesson planning tools include:

- **Recent research suggests AI EdTech tools may save teachers time when planning lessons.** An EEF trial of the impact of ChatGPT in lesson planning for KS3 science teachers found that using ChatGPT saved teachers on average 25.3 minutes per week.
- **DfE survey data suggests teachers spend a large amount of time planning and preparing lessons.** Results from the 2025, DfE Working lives of teachers and leaders survey showed that 41% of primary and secondary school teachers thought they spent "too much time" on lesson planning.
- **Trials into the impact of some EdTech lesson planning tools.** An RCT of a lesson planning tool using an AI lesson assistant, named Aila, is currently taking place in schools. The study aims to understand the impact of using Aila on time spent by teachers on lesson preparation, and to assess the quality of lesson resources produced.

Marking and feedback

AI-powered EdTech tools however are automating these routine aspects of the marking process while aiming to provide useful feedback.

Key considerations for marking and feedback tools include:

- **International research suggests potential time savings, though results should be interpreted cautiously.** A 2024 study in Indian colleges found that AI EdTech tools saved teachers 7.1 hours on grading and 2.3 hours on student feedback per week. Although promising, these substantial time savings warrant careful scrutiny, as they may not translate directly to the UK education environment due to different curriculum demands, assessment requirements, and regulatory frameworks.
- **DfE survey data suggests teachers spend a large amount of time marking work.** Results from the 2025 , DfE Working lives of teachers and leaders survey showed that 38% of primary and secondary school teachers thought they spent "too much time" marking student's work.
- **A 12-month pilot investigating how AI can support marking and feedback starting in tertiary education**. This year-long pilot, run by JISC, will test both specialist AI tools and general AI tools to reduce staff workload.

Administrative tasks

AI-powered tools are targeting inefficiencies through automation and integration by streamlining communication workflows and reducing manual reporting burdens.

Key considerations for administrative task-related tools include:

- **Pilot trials that suggest AI chatbots can save staff time on administration activities tools.** For example, Ada is a general-purpose AI chatbot developed by Bolton College to support student services by answering common queries.
- **DfE survey data suggests that most teachers are spending a large amount of time on administrative tasks**. Results from the 2025 DfE Working lives of teachers and leaders survey showed that 71% of primary and secondary school teachers thought they spent "too much time" on general administrative work.
- **A pilot investigating how AI can enhance student support services is ongoing.** LearnWise is an AI-driven platform which enhances student access to support services, which was recently piloted in fifteen colleges and universities.

Balancing evidence potential and feasibility

Testing marking and feedback tools in initial studies offers the best balance of evidence potential and feasibility and therefore present a strong fit for this pilot. The reasons for this are:

- **Evidence suggests feedback, marking and assessment tasks are the second highest priority for reducing teacher workload, after behaviour management.** As found in a 2023 EEF report, where 81% respondents identified this as a high or medium priority.
- **Teacher demand for change in marking and feedback is strong.** In a 2025 Teacher Tapp survey, both primary and secondary school teachers (22% and 27%, respectively) identified 'marking students' work' as the one job they would most like to see reduced in its time burden. This suggests there is likely to be a significant cohort of educational staff that are willing to trial these tools and there is a clear demand for new methods of minimising the burden of marking and assessment, which could help increase educator engagement for the pilot phase.
- **There is an evidence gap on the impact of AI tools on marking and feedback.** Although some studies for marking and feedback are in the pipeline, there is currently a clear need for additional evidence in these areas and new evidence in marking and feedback and administration tools would be useful to plug the evidence gap. Aligning efforts with existing research needs and currently ongoing

trials will help strengthen the overall evidence base for what works in reducing teacher workload.

- **Focusing on these tasks aligns with government innovation support for the sector.** The UK government has invested £2 million to support the development of 14 AI tools in this area. Aligning the pilot with this investment can help support evidence-informed scaling of these innovations through the testbed pipeline.

Trade-offs and risks

Focusing on marking and feedback tools involves trade-offs and risks.

While this focus is justified for Test and Learn studies, there are important limitations and risks to be acknowledged:

- **Workload pressures on non-teaching staff would not be investigated.** Focusing exclusively on marking and feedback does not build evidence for tools that support administrative and support staff members, who also face significant workload pressures. This may therefore limit the overall impact of the intervention, given the DfE's broader aim to reduce workload across all school staff. However, future studies following the pilot can address tools for non-teaching staff.
- **Concerns remain around fairness and accuracy of tools.** While these technologies could improve efficiency, there are concerns around their accuracy and ability to interpret student responses and provide equitable feedback across diverse learner groups. If not carefully implemented, such tools risk disadvantaging certain cohorts of students. Therefore, ensuring that these tools do not reinforce bias or reduce the quality of feedback is essential. It may also be necessary to implement teacher oversight or restrict tool use to non-grade-bearing assessments.
- **Workload savings may be obscured by new working practices.** Experts argued that teachers need training in how to implement AI tools effectively, which may lead to them feeling overwhelmed. As such, a period of familiarisation may be needed for teachers before the full extent of workload reduction is clear. Other research emphasises that AI tools should complement professional judgement, rather than replace it.

Offsetting risks

Future studies may look to expand the selection of tools. Limiting tools for the initial study to only those that help with marking and feedback does not mean that other tools should be discounted in subsequent studies.

However, it is important to note that any restrictions placed on tool focus will be primarily dictated by the number and type of tools available and edtech companies willing to participate. As such, remaining flexible and adapting the tool focus of any study to match the tools available will be necessary.

How to measure impact when testing tools

When intending to run evaluations of the impact of tools, there needs to be careful consideration around how to measure these impacts. This section considers how to measure the impact of tools when testing tools across all three priority areas.

Teacher workload outcomes

To evaluate the effectiveness of EdTech tools in reducing teacher and administrative workload, it is essential to identify outcomes that can be measured clearly and feasibly within the constraints of a short study. Teacher workload can be assessed in two primary ways, both using survey data:

- **Total working hours.** How many hours they spend working in total in any given period (e.g. a week).
- **Time spent on specific tasks.** How many hours they spend on specific tasks, such as marking, providing feedback or lesson planning, in any given period.

Capturing the time taken for specific tasks is particularly important. This is because it's possible for an EdTech tool to successfully reduce time spent on one task (e.g. marking) while increasing the time spent on another task (e.g. lesson planning). If only total working time was measured this nuance would be missed.

It's also important to measure whether a tool inadvertently increases a teacher's workload even while saving time in one area. For example, an automated grading tool might reduce marking time, but if it fails to provide the detailed feedback data teachers need for lesson planning, any time savings could be offset by the extra work required to gather that information elsewhere. In addition to measuring time taken for a task, the impact of EdTech tools on teacher workload can also be measured through related outcomes. These outcomes could also be impacted in several ways by the introduction of EdTech tools. They include:

- **Wellbeing**. A tool may help reduce the working hours of a teacher and therefore reduce stress and improve wellbeing.
- **Job satisfaction**. Even if working time does not fall overall, teachers may be able to make more efficient use of their time, which could lead to higher satisfaction with their work.

Another relevant dimension to consider is the quality of the output produced using EdTech. In the case of marking, it is important to assess whether the EdTech tool marks work correctly. This is especially relevant in subjects where answers can be clearly classified as right or wrong. A tool might not reduce the total time spent marking (for

example, if teachers need to spend significant time reviewing the tool's outputs), but it could still deliver benefits by enabling more detailed or personalised feedback to learners.

In sum, while time saved is one important indicator of teacher workload, it may not need be the sole focus. Evaluations can also consider how that time is used, the teacher experience, and the quality of educational delivery supported by the tool. To ensure a full and accurate picture is gained on the impact of a time-saving tool, the delivery partner may want to work closely with the evaluation team so that these indicators of teacher workload are being measured. This will allow the pilot's studies to be well designed and produce results that capture the nuanced impact of these tools, allowing the DfE to consider their impact on teacher's work in detail.

Pupil and student learning outcomes

Measures of student attainment are in some ways simpler than measuring teacher workload, as assessments are carried out at the end of different key stages and for colleges during different stages of learners' courses. However, this would limit the year groups that the initial study could examine to those taking exams in the year of the evaluation and mean that the time horizon on the evaluation would lengthen.

In assessing attainment in this way, the evaluation would be judging longer term learning. For assessment of short-term learning (or evaluations with a tighter timeframe), it might be feasible to conduct in class tests for the subjects of interest. It is common practice for schools to conduct routine testing through the year (outside of what they report to the DfE as part of the National Pupil Database). These measures would be less burdensome for schools to collect (as they would be doing so anyway) and so preferable to bespoke tests.

Inclusion of all children

Measures of inclusion would likely be assessed through survey questions. This can involve asking participants direct questions about how included they feel or using more indirect indicators such as how connected they feel to other learners, how happy they are at school or their sense of belonging.

Attendance data may also be useful as a proxy for engagement with learning. Similarly, surveys of parents or teachers asking similar questions could be conducted to gain a broader perspective on inclusion.

Evaluation rationale and options

This section discusses the evaluation methods that could be used to assess both Test and Learn or Evidence Hub based studies, should the DfE choose these testbed models for the Pilot. It begins by outlining the rationale for conducting an evaluation. It then describes the different evaluation approaches that may be needed for each study.

The DfE's three core objectives for this project are: identifying EdTech tools, generating evidence and testing the testbed. To meet these goals, any evaluation will need to focus on the EdTech tools implemented within the structured environment of the testbed, examining both their impact and the context of their delivery.

Specifically, an evaluation will:

- **Explore implementation and delivery.** The evaluation will explore how studies within the testbed are implemented and delivered. For example, exploring how each tool was deployed and experienced by educators and learners within the operational framework of the testbed, including any enabling or constraining factors. This directly addresses the DfE's objective of testing the testbed.
- **Generate insights into the effectiveness of the tools on key outcomes**. The evaluation will explore the effect each tool has on addressing a specific priority outcome. For example, the number of hours saved by staff on a specific task or increases in student test performance. This directly addresses the DfE's objective regarding generating evidence of impact.

This dual focus, on both implementation and outcomes, would help build a comprehensive understanding of not just what works, but how and why it works in real-world settings. Additionally, evaluating the tools would also help understand their scalability, helping determine whether the tools can succeed across different school and college contexts, and for a diverse range of learners.

Comparing different evaluation approaches

The overall scope and depth of evaluations in will depend on the methodological approach adopted.

This section outlines these options, which are summarised in Table 2. A summary of a typical evaluation progress, and the key stages, is given in Annex 1.

Table 2: A summary of the evaluation approaches that could be used the EdTech Impact Testbed

	Test and Learn approach	RCT or MCD
Evaluation methods	Pilot evaluation A lighter-touch descriptive evaluation focused on implementation, uptake, and perceived value of EdTech tools in a small number of settings.	(a) Implementation and process evaluation Qual and quant analysis of how EdTech tools are delivered, used and experienced in real settings. Focus on fidelity, acceptability, reach and context (b) Impact evaluation Robust causal design using either a randomised controlled trial (RCT) or a matched comparison design (MCD). Focus on quantifying effects on priority outcomes (e.g. reducing workload)

Evaluating Test and Learn based studies

A pilot evaluation does not aim to establish causal impact. Instead, it focuses on assessing the feasibility of implementation, examining early-stage promise, and refining both the tools and future evaluation design. This type of evaluation is particularly suited to early-stage innovations and interventions that have not yet been widely tested in real-world education settings. EEF guidance, published in 2023, discusses the major considerations for designing a pilot evaluation.

A well-conducted pilot evaluation provides a structured and evidence-informed way to answer the following core questions:

1. Can the EdTech tool be delivered as intended in the real-world settings of schools and colleges?
2. What adaptations or support are needed to ensure successful implementation?
3. Does the tool show enough promise on short-term outcomes to warrant investment in a full-scale evaluation?

Pilot evaluations typically involve a smaller group of schools or colleges (around 25) and do not require a control group. This makes them well-suited to contexts where budgets, timelines, or sample sizes limit the feasibility of a randomised or quasi-experimental design. While they do not produce robust causal evidence, they can still provide valuable

early insights into the functionality of the testbed model and the implementation readiness of the tools under consideration.

A descriptive, mixed-methods approach is typically used to generate a comprehensive picture of how tools operate in practice. Data sources can include:

- **Qualitative interviews and focus groups.** School and college leaders, teachers, and administrators discuss perceived value, usability, and barriers to adoption.
- **Tool-generated usage data.** Can be used to assess levels of engagement and patterns of use across different contexts.
- **Surveys**. Can capture self-reported outcomes such as changes in workload, satisfaction, or teaching practices.
- **Administrative data**. Where available, this data can be used to examine associations between tool use and selected indicators.
- **In-depth case studies.** Selected schools, colleges or user groups can be used to highlight contextual factors that shape implementation.

A key strength of pilot evaluations is their flexibility. They can accommodate smaller sample sizes, proceed without control groups, and adapt more easily to changing delivery conditions. However, this also limits their ability to assess impact with confidence. Any observed changes cannot be definitively attributed to the tool, as there is no counterfactual comparison.

To maximise learning, pilot evaluations should pre-specify clear criteria for success. These might include minimum usage thresholds, positive engagement from teachers, or measurable shifts in perceived workload. Establishing these indicators in advance ensures that the evaluation generates actionable conclusions and supports decisions about whether the tool is ready for more rigorous testing. It will also support make evidence-based decisions about which tools to prioritise for further investment.

Pilot evaluations can often be completed over a single academic term, which makes them particularly useful when delivery timelines are tight. However, strong implementation discipline is still essential. Even in non-causal studies, inconsistent delivery can limit learning. As such, clear protocols and monitoring mechanisms may want to be built into the design.

While the absence of causal inference is a limitation, pilot evaluations can play a crucial role in laying the groundwork for future studies held within the testbed pilot. They help identify delivery challenges, refine outcome definitions, and inform sampling and power assumptions for subsequent RCTs or quasi-experimental designs. In this way, pilot evaluations are not a substitute for RCTs, but a complementary approach within the

testbed's wider evidence pipeline, especially where the goal is to test, learn, and adapt at pace.

Evaluating Evidence Hub based studies

This more rigorous evaluation involves (a) an impact evaluation (IE) combined with (b) an implementation and process evaluation (IPE). The IE establishes the causal relationship between EdTech and outcomes within the chosen priority area for analysis, whilst the IPE focuses on drawing insights from how the intervention was delivered.

Impact evaluations

To assess whether the EdTech tools are driving improvements in the selected priority outcome (such as reducing teacher workload), the evaluation must adopt a design capable of establishing causality. There are two main options for doing this: randomised controlled trials or quasi-experimental designs (of which there are multiple types).

Each design carries different trade-offs in terms of data collection requirements, the strength of evidence produced, and budget implications. These should be carefully considered against the study's goals, time constraints, and available resources. A full discussion of RCTs and quasi-experimental designs (such as Matched Comparison Design studies) can be found in Annex 3.

Implementation and process evaluations

The purpose of an IPE is to understand how and why an intervention works (or does not), for whom, and under what circumstances. IPEs are especially useful for complex, real-world interventions where multiple factors influence success. They provide essential context that helps explain the findings of an impact evaluation, revealing insights that go beyond headline quantitative results.

An IPE typically explores several key dimensions:

- **Fidelity.** Was the intervention delivered as intended?
- **Exposure and reach**. How much was the intervention used? What proportion of those who were supposed to use it did?
- **Acceptability**. How well did stakeholders engage with the intervention?
- **Mechanisms.** Why might the intervention result in change? Are there particular channels that are more likely than others?
- **For whom and in what context.** Does the tool work for some schools or colleges and not others? What are the reasons for this? How well might the intervention scale up outside of the trial as a result?

As part of the IPE, participants will take part in interviews and focus groups. The IPE may also include case studies, focusing on specific schools and colleges or tools (depending on the chosen unit of analysis). These case studies will allow for a more in-depth exploration and can be compared to highlight variations in implementation or contextual differences. Quantitative elements can also be introduced, such as back-end user data from EdTech platforms or surveys, to gain a wider range of insights.

However, IPEs are relatively resource intensive. Interviewing even modest numbers of participants takes time, both for the interviews themselves and subsequent transcription and analysis. Additionally, extra resources could be allocated to the recruitment of participants, as participants can often be difficult to engage.

Annexes

Annex 1: A typical evaluation cycle has three core phases

Given the short evaluation window for test and learn studies (approximately 6-12-weeks), careful planning is required to ensure that the evaluation generates meaningful and credible insights.

A key consideration within this period is the need to train participating teachers to use the EdTech tools effectively. This training may take place prior to the formal start of the evaluation or during its early stages, depending on scheduling and readiness. Ensuring that teachers are confident and capable in using the tool is essential to avoid skewing results due to implementation issues.

A typical evaluation cycle includes three core stages:

1. **Set-up and design.** This includes finalising the evaluation design, agreeing on roles and responsibilities, selecting participants, and delivering key documentation such as an evaluation protocol and a Statistical Analysis Plan (SAP) for randomised trials. Importantly, this stage also involves preparing training materials and onboarding schools, colleges and EdTech providers.
2. **Data collection.** This period includes the tool's active use in schools and colleges, alongside baseline and follow-up data collection. Depending on the intervention, a brief 'ramp-up' phase may be needed to allow teachers to become familiar with the tool.
3. **Analysis and reporting.** Following the end of data collection, the evaluation team will analyse the findings and deliver a final report, drawing out lessons on the tool's impact, implementation fidelity, and practical recommendations for DfE and delivery partners.

These stages may overlap slightly, particularly where training and baseline data collection coincide. However, ensuring that sufficient time is allocated for setup and analysis outside of the core implementation window is essential. Compressing these surrounding phases could compromise either the quality of the evaluation design or the rigour of the conclusions drawn from the data.

Annex 2: Evaluation considerations for RCTs and MCDs

This section gives a summary of the differences between RCTs and MCDs, including how they could be implemented within the EdTech Impact Testbed pilot.

Randomised Controlled Trial (RCT)

An RCT represents the most rigorous method for establishing causal impact. In an RCT some units (schools/colleges, year groups, or classes) would be randomly assigned to either receive treatment (in this case the EdTech tool) or to act as a control. This overall sample would be assigned from schools and colleges that have put themselves forward to be involved in the trial. The fact that the assignment process for treatment is random means that background factors (such as a school's/ college's motivation to use EdTech tools or their teacher and learner's levels of tech literacy), which can be difficult to measure accurately, will not drive any identified effects.

Alongside any measures of workload that are available for all schools and colleges in administrative data, surveys can be collected in both treatment and control groups. This allows the impact to be assessed on a wider range of outcome measures; examples of which are highlighted in the section above.

Randomisation can be done at several different levels:

- **Education setting level.** Some schools and colleges would be randomised to treatment and others would not. This approach has two potential risks:
    - **Education providers may be reluctant to participate due to added burden.** For example, schools and colleges may be expected to carry out additional tasks, such as administering surveys. However, this concern can be mitigated by the tool being provided once the data collection period is over.
    - **Risk of tool use by control group.** If the chosen EdTech tool is closer to the consumer-end of the market (e.g. such as trials of lesson planning with ChatGPT), it is possible that some teachers in schools and colleges in the control group end up using it anyway. This would bias effects downwards.
- **Teacher/class/year group level**. Alternatively, some units within a school or college could be assigned to treatment, whilst others are not. Here, there is the potential for 'spillover effects', where the impact of the EdTech tool spills over onto control units and biases the estimated impact of the tool. There could also be direct use of the tool by the control group either by accident or because it is perceived to be effective.

The higher the level of randomisation (where school/college level is “higher” than class or year group level) the larger the sample size required to ensure sufficient statistical power. Additionally, education providers or individuals assigned to control groups would need to refrain from using other AI tools during the study period, which may both be frustrating and mean that monitoring is required of “business-as-usual” practices to ensure fidelity to the study design.

While this approach offers the strongest level of evidence, it also demands tight implementation discipline and careful planning to ensure the trial is sufficiently powered to detect meaningful effects.

Furthermore, to determine whether any of the estimated effects are unlikely to have happened by chance, the RCT will need to have enough statistical power. Formal power calculations would be required to be able to say what sample size is needed for the study to be sufficiently powered, but a rule of thumb would be that each treatment arm (including the control group) would need 50-70 schools and colleges. Any less than this sample size and the DfE will be left in a position where they are unable to say with any certainty whether particular tools had an impact. Power calculations can be conducted by the evaluation partner at the point of designing the evaluation protocol and statistical analysis plan.

With an estimated sample size of 100-200 schools and colleges feasible with an approximate £1 million budget, this approach can realistically support testing just one tool. Splitting the sample across multiple treatment arms would dilute statistical power, making it difficult to detect reliable effects.

Matched Comparison Design (MCD)

Without random assignment, it is still possible to approach an evaluation in a way that still leads to robust estimates, namely through a quasi-experimental design (QED). One potential QED in this context is a matched comparison design (MCD).

A major advantage of this approach is logistical simplicity: all interested education providers receive the intervention without requiring them to be randomly allocated to either treatment or control groups. Comparisons are then made using a control group of schools and colleges, which are similar to the treatment schools and colleges, but are not trialling the products. These comparison schools and colleges would be selected from administrative data (held by the DfE) based on their observed characteristics such as size, location, proportion of students eligible for FSM size, gender balance of the learners, or deprivation levels in the local area.

However, this approach has two key limitations:

- **Risk of unobserved differences biasing results.** Matching can only be based on observable variables and cannot account for unobserved differences such as school or college culture, leadership, or parental preferences which randomisation (in the RCT method) would account for. These unmeasured factors can introduce selection bias, limiting the validity of the findings and the ability to draw firm causal inferences.
- **Limited outcome measures.** As matched comparison designs rely on administrative data for the control group, the evaluation is restricted to the outcomes available in those datasets. This excludes any outcomes that require direct data collection, such as teacher surveys or feedback, narrowing the scope of the analysis. Additionally, access and processing times for administrative data can delay when results become available.

With 100–200 schools and colleges in the initial study, more than one tool could be tested. This is because matched comparison designs do not require large, powered samples for each treatment arm. Instead, they allow the intervention group to be split across multiple tools without compromising the ability to construct suitable matched comparators from administrative data.

Annex 3: Types of evidence

This section provides a non-exhaustive list of the types of evidence that can be used to measure the potential impact of EdTech tools. Each type of evidence lists at least one challenge with collecting evidence in the classroom.

- **Randomised controlled trials (RCTs).** Viewed as the gold standard for measuring causal impact. They tend to be an expensive form of evidence gathering, disruptive to participants as they require strict adherence to a protocol. There are also possible ethical concerns when randomising participants in an education setting, as some learner would not receive the intervention (in this case the tool)
- **Quasi-experimental studies.** Product evaluations and comparisons without full randomisation.
- **Pre-post assessments.** Measurements for comparison of outcome gains over time. Learning improvements may be influenced by external school/college factors that can’t be controlled for in the study.
- **Rapid test cycles.** Iterative testing with feedback loops between educators and developers. Requires a teacher’s buy-in. Rapid cycles may clash with academic year timetables.

- **Product pilots.** Short-term trials to assess usability and fit. Schools and colleges may lack capacity for meaningful pilot integration due to resource and expertise constraints.
- **Survey responses.** Structured feedback on perceptions and experiences. Low response rates often a risk, as well as respondents providing socially desirable answers.
- **Usage analytics.** Quantitative data on tool usage. Data can be incomplete or inconsistent as schools and colleges vary in the quality of their tracking systems.
- **Observational data.** Notes or logs from classroom observations. Can be time-consuming for staff. Observer presence may also influence behaviour. In the classroom this may be teachers influencing students.
- **User reviews.** Structured feedback from users. Samples can be limited if not well designed.
- **Anecdotal reports.** Individual reviews of tools, which have been used in a specific setting. Can be subjective to the teacher's viewpoint and may depend on their enthusiasm.

Annex 4: Bibliography

- Batty, R., Florescu, A., Wong, A. and Sharples, M. (2019). *EdTech testbeds: Models for improving evidence.*
- Roschelle, J., Means, B.M. and Shear, L. (2015). *Using technology and evidence to promote cultures of educational innovation: The example of science and mathematics education.*
- Cukurova, M., Luckin, R. and Clark-Wilson, A. (2019). *Creating the golden triangle of evidence-informed education technology with EDUCATE. British Journal of Educational Technology.*
- Stringer, E., Lewin, C. and Coleman, R. (2019). *Using digital technology to improve learning: Guidance report.* Education Endowment Foundation.
- Scanlon, E., Sharples, M., Fenton-O'Creevy, M., Fleck, J., Cooban, C., Ferguson, R., Cross, S. and Waterhouse, P. (2013). *Beyond prototypes: Enabling innovation in technology-enhanced learning.*
- Vanbecelaere, S., Adam, T., Sieber, C., Clark-Wilson, A., Adorno, K.B. and Haßler, B. (2023). *Towards systemic EdTech testbeds: A global perspective.*
- Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016). *Implementation and process evaluation (IPE) for interventions in educational settings: An introductory handbook.*
- Takala, H., Kuo, T-L., Duysak, E., Bhatti, S., Stoilova, E., Fletcher, A., McGuinness, N., McKaskill, M. and Fugard, A. (2025). *Stop and think: Learning counterintuitive concepts – Evaluation report.* Education Endowment Foundation.
- What Worked Education. *Guide to micro-randomised controlled trials (mRCTs).*

- Global EdTech Trialing Network, Adorno, K. and Mote, E. (2023). *Tenets and principles of EdTech trialing networks and environments within the US.*
- NHS England. *Test Beds Programme: Information governance learning from Wave 1.*
- Sacks, I. (2025). *Stanford initiative helps scale what works in education.*
- Heffernan, N.T. and Heffernan, C.L. (2014). *The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching. International Journal of Artificial Intelligence in Education.* doi:10.1007/s40593-014-0024-x
- Prihar, S. et al. (2022). *Exploring common trends in online educational experiments.* doi:10.5281/zenodo.6853041
- Roy, P., Poet, H., Staunton, R., Aston, K. and Thomas, D. (2024). *ChatGPT in lesson preparation – A teacher choices trial.* Education Endowment Foundation.
- Carter, A. (2020). *Sandboxes: Testing the strategy in Malawi.* EdTech Hub.
- Education Endowment Foundation (2023). *Guidance for pilot evaluations.*
- Vojtkova, M., Smith, L. and Vadgama, S. *Test and learn: A playbook for mission-driven government.*
- Arntzen, S., Wilcox, Z., Lee, N., Hadfield, C. and Rae, J. (2019). *Testing innovation in the real world.*
- European EdTech Alliance. *Exploring the EdTech testbed ecosystem.*
- EmpowerEd Project (2024). *The European EdTech ecosystem roadmap.*
- Dijkstra, P. (2022). *Under the bonnet: A review of EdTech testbeds in Europe – Examples of cooperation between EdTech developers and the education community.*
- UCL Institute of Education. *Global approaches for the systemic piloting and trialling of school educational technologies.*
- Åkerfeldt, A. (2024). *Nordic EdTech/Digitalisation: Generating evidence through piloting testbeds.*
- [No author]. *Qatar's EdTech Testbed: Leveraging global insights for effective EdTech testbed design in Qatar.*
- Department for Education (2022). *Implementation of education technology in schools and colleges.*
- Taggart, S. and Roulston, S. (2025). *Trailblazing NI genAI in education: A proof-of-concept study in schools in Northern Ireland with MS Copilot.* Ulster University.
- Fishman, B., Penuel, W., Allen, A., Cheng, B. and Sabelli, N. (2013). *Design-based implementation research: An emerging model for transforming the relationship of research and practice.*
- Edovald, T. and Nevill, C. (2020). *Working out what works: The case of the Education Endowment Foundation in England.*
- EdSafe AI (2025). *Opportunity at scale: The case for public infrastructure for AI in education.*
- The EdTech Genome Project (2021). *The EdTech Genome Project report.*

- Jenkinson, J. (2009). *Measuring the effectiveness of educational technology.*
- Lee, S.S. (2022). *Qatar's EdTech Testbed: Building collaborative partnerships for innovative teaching and learning outcomes.*

Annex 5: Glossary

- **EdTech (Education Technology):** EdTech refers to the use of digital tools and resources to improve teaching, learning and school administration and management. It encompasses a wide range of technologies, including hardware, software, services and AI-based tools.
- **Testbed:** A structured environment in which EdTech tools are trialled in real-world educational settings to evaluate their impact.
- **Pilot Testbed:** The initial, small-scale implementation of an EdTech testbed used to test feasibility and generate early insights before scaling.
- **Priority Areas:** The DfE's three focus areas for EdTech testing: teacher/administrative workload, learner learning outcomes, and inclusion of all students, including those with SEND.
- **Teacher Workload:** The time and effort teachers spend on non-teaching tasks, such as lesson planning, marking and administration.
- **Learning Outcomes:** Measures of learner progress and attainment, often assessed through exams, tests, or ongoing classroom data.
- **SEND (Special Educational Needs and Disabilities):** Refers to children and young people requiring additional support to access education; improving inclusion for SEND learners is a DfE priority.
- **Expression of Interest (EoI):** An open call for schools, colleges, and EdTech companies to apply to participate in the testbed pilot.
- **Co-design Model:** A testbed approach where educators, learners, and EdTech developers collaboratively design prototypes and research studies.
- **Test and Learn Model:** A testbed model focused on rapid-cycle, real-world testing of mid-stage tools to generate early insights and promise of impact.
- **Evidence Hub Model:** A testbed model emphasising formal evaluations (e.g., RCTs) of mature tools to generate robust, generalisable evidence.
- **EdTech Network Model:** A testbed model that prioritises knowledge-sharing and collaboration across educators, policymakers, and developers.
- **Randomised Controlled Trial (RCT):** A rigorous impact evaluation design where education settings are randomly assigned to either receive or not receive an EdTech tool.
- **Matched Comparison Design (MCD):** A quasi-experimental evaluation method comparing education settings using an EdTech tool with similar schools and colleges not using it.
- **Pilot Evaluation:** An exploratory assessment focused on feasibility, usability, and early promise rather than causal impact.

- **Impact Evaluation:** A study designed to assess whether an EdTech tool directly causes improvements in outcomes such as workload or attainment.
- **Implementation and Process Evaluation (IPE):** An evaluation approach examining how a tool was used, for whom it worked, and why, alongside barriers and enablers.

Annex 6: Typical evidence outputs for each testbed model

Table 3: Typical evidence outputs for each testbed model

Testbed model	Typical evidence outputs	Examples of testbeds and outputs
Co-Design	Design briefs and prototypes User journey maps Qualitative insight reports School-level case studies	**LeanLab Education (USA):** - Research reports on usability and promise - Case studies with teacher narratives - Feedback dashboards for vendors
Test and Learn	Pilot evaluation reports (mixed-methods) Implementation notes Teacher/staff workload surveys Short thematic case studies	**National Education Lab AI (Netherlands):** Short pilot reports on AI in classrooms
Evidence Hub	RCT or quasi-experimental reports Implementation and process evaluations Meta-analyses and synthesis reports Technical notes on methods	**Education Endowment Foundation (UK):** RCT evaluation reports + plain-language summaries
EdTech Network	Practice guides and toolkits Infographics and one-page summaries Webinars and podcasts School leader briefings	**EdTech Hub (Global):** Toolkits, evidence syntheses, dissemination via online platforms

Department for Education

ISBN: 978-1-83870-737-8

For any enquiries regarding this publication, contact www.gov.uk/contact-dfe.

This document is available for download at www.gov.uk/government/publications.