

The impact of intelligent tutoring systems in general education

A systematic review of causal evidence on learning outcomes, motivation, and socio-emotional effects

Please cite this publication as:

Theis, D., Thomsen, M., Vogel, C., and Wasserfuhr, V. (2026). 'The Impact of Intelligent Tutoring Systems in General Education', *ENESET report*, Luxembourg: Publications Office of the European Union. doi: 10.2766/0766181.

AUTHORS OF THE REPORT:

- **Dominik Michael Theis**, Institute for Innovation and Technology (iit)
- **Dr. Maren Thomsen**, Institute for Innovation and Technology (iit)
- **Dr. Cathrin Vogel**, Institute for Innovation and Technology (iit)
- **Veit Wasserfuhr**, Institute for Innovation and Technology (iit)

PEER-REVIEWERS:

- **Kristof De Witte**, KU Leuven and ENESET Scientific Coordinator
- **Miguel Arevalillo-Herráez**, University of Valencia

LANGUAGE EDITOR:

- **James Nixon**, Freelance editor

REPORT COORDINATOR:

- **Stéphanie Crêteur**, PPMI (Part of the Verian Group)

ABOUT ENESET

ENESET is an advisory network of experts specialising in the social and economic dimensions of education and training. It continues the work of previous networks coordinated for the European Commission's Directorate-General for Education and Culture. ENESET draws on the legacy of NESET and EENEE to support evidence-based policymaking through high-quality research, analysis, and policy advice. The network is coordinated by the Public Policy and Management Institute (PPMI). For inquiries, please contact us at: eneset@ppmi.lt.

Contractor:

PPMi

Part of the
Verian Group

PPMI (Part of the Verian Group)
Gedimino Ave. 50, LT - 01110 Vilnius, Lithuania
Phone: +370 5 2620338 Fax: +370 5 2625410
www.ppmi.lt

Director: Egidijus Barcevičius

EUROPEAN COMMISSION

Directorate-General for Education, Youth, Sport and Culture
Directorate A — Policy Strategy, Country Intelligence and Youth
Unit A.3 — Evidence-Based Policy and Evaluation

E-mail: eac-unite-a3@ec.europa.eu

European Commission
B-1049 Brussels

The impact of intelligent tutoring systems in general education

A systematic review of causal evidence on learning outcomes, motivation, and socio-emotional effects

Manuscript completed in September 2026

1st edition

This document has been prepared for the European Commission; however, it reflects the views only of the authors, and the European Commission is not liable for any consequence stemming from the reuse of this publication.

Luxembourg: Publications Office of the European Union, 2026

© European Union, 2026



The reuse policy of European Commission documents is implemented by Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39). Unless otherwise noted, the reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence (<https://creativecommons.org/licenses/by/4.0/>). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

For any use or reproduction of elements that are not owned by the European Union, permission may need to be sought directly from the respective rightholders.

Print ISBN 978-92-68-39045-0
PDF ISBN 978-92-68-39044-3

doi:10.2766/0199114
doi:10.2766/0766181

NC-01-26-113-EN-C
NC-01-26-113-EN-N

Contents

1. Executive summary	7
2. Introduction	10
2.1. Defining intelligent tutoring systems	11
2.2. Conceptual impact framework	12
2.3. Purpose of this review	16
2.4. Research question.....	16
2.5. Contribution	17
3. Methods	19
3.1. Review design and protocol development	19
3.2. Search strategy.....	19
3.3. Eligibility criteria	19
3.4. Screening procedure	20
3.5. Data extraction.....	21
3.6. Quality appraisal.....	21
3.7. Synthesis	21
3.8. Adapted PRISMA flow diagram	22
3.9. Working papers.....	23
4. Results	24
4.1. Descriptive overview of the reviewed studies	24
4.1.1. Study design.....	24
4.1.2. Countries where research was conducted	26
4.1.3. Timing of research	28
4.1.4. School level and age group.....	29
4.1.5. Subject areas covered	30
4.1.6. Outcomes examined	31
4.2. Definition and types of ITSs	31
4.2.1. Operationalisation of ITSs in the studies reviewed	31
4.2.2. Core components of ITSs in the studies reviewed	33
4.2.3. Features of ITSs in the studies reviewed	34
4.2.4. Types of ITSs in the studies reviewed.....	37

4.3.	Main findings.....	39
4.3.1.	Academic performance and skill development	39
4.3.2.	Motivation and engagement.....	43
4.3.3.	Well-being and other socio-emotional effects.....	46
4.4.	Moderators and mechanisms	47
4.4.1.	Implementation contexts	48
4.4.2.	Feedback.....	49
4.4.3.	Adaptive sequencing.....	50
4.4.4.	Scaffolding.....	53
4.4.5.	Technological design features	55
4.4.6.	Educational stage and age group	57
4.4.7.	Subject domain	60
4.4.8.	Role of teachers and parental involvement	65
5.	Discussion.....	69
5.1.	Summary of key findings	69
5.2.	Synthesised impact framework.....	72
5.3.	Limitations of this review.....	77
5.4.	Conclusions	80
6.	Implications and recommendations.....	84
6.1.	Policy-relevant takeaways	84
6.1.1.	Specify the intended use and observed effects before selecting an ITS..	85
6.1.2.	Translate pedagogical mechanisms into quality and procurement requirements.....	85
6.1.3.	Fund implementation capacity as part of the intervention.....	86
6.1.4.	Pilot under routine conditions and make scaling conditional on evidence	87
6.1.5.	Build a shared European evidence and coordination framework.....	88
6.2.	Recommendations for national and regional policymakers.....	89
6.2.1.	Intended use and strategic alignment	89
6.2.2.	Funding, procurement and quality requirements	89
6.2.3.	Implementation capacity and professional development	90
6.2.4.	Implementation, equity and monitoring	91
6.3.	Recommendations for schools in the European Union	91
6.3.1.	Intended use and system selection	92
6.3.2.	Teacher roles and implementation conditions	92
6.3.3.	Pilot, monitoring and adaptation.....	93
6.3.4.	Transferability, equity and safeguards	93
6.4.	Recommendations for European Union institutions	94
6.5.	Implications for evaluating emerging generative AI-based tutoring.....	95
7.	References.....	97

1. Executive summary

Intelligent tutoring systems (ITSs) can contribute meaningfully to pupils' learning in general education, but their benefits are neither automatic nor uniform. The evidence reviewed is most extensive and consistently positive for learning outcomes and skill development. Findings on motivation and engagement are generally favourable but more heterogeneous, while evidence on well-being and socio-emotional outcomes is smaller and indicative rather than conclusive. Across these domains, educational value depends on how systems translate learner information into appropriate instructional support and how their use is embedded in teaching and learning.

Purpose, scope and approach

This report synthesises evidence on ITSs and related AI-based adaptive learning tools in primary, lower-secondary and upper-secondary education. It examines their effects on learning outcomes, motivation and engagement, and socio-emotional development, alongside the pedagogical mechanisms and implementation conditions associated with these effects. ITSs are understood as systems that combine learner information, subject-domain structures and pedagogical decision-making to adapt instruction. Functionally equivalent adaptive learning tools are included even where studies do not explicitly describe the canonical ITS architecture. Systems based on generative AI and large language models are excluded from the review's operational definition.

The review includes **65 studies**, identified through a structured search targeting English-language evidence from 2020–2026. The search relied primarily on Semantic Scholar via Elicit. Following automated prioritisation of 9,985 records, two reviewers manually screened 400 abstracts and assessed 94 full texts. Included studies underwent structured data extraction and quality appraisal. The evidence is predominantly intervention-oriented, with quasi-experimental studies and randomised controlled trials forming the largest groups. Findings are synthesised narratively and thematically rather than through a meta-analysis; the report therefore does not provide a pooled estimate of effectiveness.

Main findings

Learning outcomes and skill development show the strongest evidence of benefit. Studies frequently report improvements in achievement, conceptual understanding, problem solving, reasoning, mastery and reading-related performance. Positive findings also extend to language learning, science and computer science, although evidence outside mathematics is thinner. Reported standardised effects range from small to very large, but their magnitude and consistency vary with study design, comparison condition, learner group, curricular content and intensity of use.

Motivation and engagement are often improved, but less consistently than learning outcomes. Reported benefits include greater enjoyment, interest, confidence, self-efficacy, participation, persistence and practice intensity. However, results differ across motivational constructs and behavioural indicators, and some measures remain unchanged or favour comparison conditions. Short interventions and first-time use also make it difficult to distinguish sustained motivational benefits from novelty effects. Engagement is therefore relevant both as an outcome and as a condition supporting continued learning.

Socio-emotional findings are favourable within a comparatively small evidence base. Studies examining these outcomes report positive affective experiences, greater confidence and perceived competence, and reductions in mathematics anxiety, frustration, mental demand and cognitive load. These findings suggest that ITSs may influence pupils' emotional experience during learning as well as their achievement.

Mechanisms and conditions for effectiveness

The review identifies **feedback, adaptive sequencing and scaffolding** as the central pedagogical mechanisms through which ITSs support learning. Feedback appears particularly relevant when it is timely, specific, explanatory and linked to errors or next steps. Adaptive sequencing adjusts task order, difficulty, pace and learning pathways to pupils' performance or estimated mastery. Scaffolding provides graduated support through hints, prompts, worked examples and guided practice. These mechanisms need to be calibrated to learners' needs: poorly timed or excessive support can encourage superficial task completion, confusion or dependence on system guidance.

Technological features such as dashboards, conversational agents, gamification and multimedia have more conditional effects. They appear most useful when they make instructional support understandable, usable or actionable—for example, by helping teachers identify learning needs or enabling pupils to monitor progress. Their presence alone does not establish effectiveness. The report's synthesised impact framework consequently distinguishes the diagnostic–adaptive layer, which interprets learner information, from core pedagogical mechanisms and the interaction features through which support is delivered.

Implementation is part of the intervention, not merely a delivery consideration. Sustained participation, sufficient opportunities for practice, curriculum alignment and reliable organisational support repeatedly shape outcomes. Teachers interpret system-generated information, connect it to curricular goals and decide when additional instruction or support is needed. Some hybrid models combining ITSs with human tutoring produce stronger outcomes than software-only use. Teacher preparation and ongoing support are therefore important implementation conditions, although their separate causal effects are rarely established.

Implications for policy and practice

For **national and regional policymakers**, the report recommends specifying the educational need, pupil group, subject, intended outcome and realistic comparison condition before funding or procuring an ITS. Quality requirements should address how learner information informs feedback, sequencing and scaffolding, alongside curriculum alignment, teacher oversight, usability and accessibility. Funding should cover implementation capacity — including preparation time, professional development, infrastructure, scheduling and ongoing support — rather than software access alone. Wider deployment should depend on evidence from pilots under conditions comparable to intended use, assessed against predefined criteria for continuation, adaptation or discontinuation.

For **school authorities and school leaders**, ITSs should be introduced as teacher-mediated instructional supports. Schools should define their curricular purpose, planned pattern of use and teachers' responsibilities, and test systems with teachers and pupils before broader implementation. Monitoring should distinguish learning outcomes from engagement and satisfaction, while also examining exposure, completion, technical interruptions, teacher workload and differences between learner groups. Implementation should be adapted, paused or discontinued where the intended educational value is not demonstrated or unintended inequalities emerge.

For **European Union institutions**, the principal contribution lies in strengthening comparable evidence and coordination while respecting Member States' responsibilities. Priorities include common evaluation and reporting principles, reusable indicators, a shared evidence and implementation repository, and comparative studies under routine European school conditions. Such work should cover successful and unsuccessful implementations, longer-term outcomes, implementation and maintenance costs, teacher-facing effects and subgroup differences. Recurring pedagogical mechanisms and implementation conditions offer more transferable points of reference than particular products or reported effect sizes.

2. Introduction

Artificial intelligence (AI) is becoming increasingly relevant in school education, both as a topic for policy debate and as a growing set of tools intended to support teaching and learning. Within this broader landscape, intelligent tutoring systems (ITSs) and related AI-based adaptive learning tools are of particular interest because they represent comparatively well established forms of educational AI. Unlike more recent applications that are based on large language models (LLMs), these systems are rooted in long-established forms of AI used in education, including rule-based and model-based approaches to adaptive instruction. They are typically based on architectures that combine learner models, domain models and pedagogical rules, and are associated with a better-developed base of empirical evidence (Ma et al., 2014). Historically, some of these systems would have been discussed using labels such as “computer-assisted instruction” or “intelligent computer-aided instruction”. In this report, however, the systematic use of learner data in conjunction with pedagogical or domain models to adapt instructional support serves as an operational criterion for their inclusion within the scope of a review of AI-based systems. This criterion describes the form of AI relevant to the review; it is not intended as a defining technical criterion for AI itself.

Systems based on generative AI and LLMs are not classified here as ITSs. The present review focuses on the classical ITS tradition of structured learner-adaptive architectures (Nwana, 1990; Holmes et al., 2019), in which instructional adaptation is organised through learner, domain and pedagogical models. Systems that primarily centre on probabilistic content generation fall outside this operational definition. This boundary is methodological rather than substantive: although generative and hybrid AI may increasingly be integrated into future ITS architectures, the present review focuses on those systems for which causal evidence in school settings is sufficiently well established. This makes ITS a particularly relevant object of analysis for an evidence-informed discussion of AI in general education. This scope was agreed with DG EAC at the outset, and reflects the report’s role within a broader analytical package on AI in education. Generative AI, LLMs and related recent developments are addressed in companion reports.

At the same time, the growing visibility of AI in education has not been matched by equally robust pedagogical guidance on its educational value. Public and policy debates often oscillate between optimistic assumptions that AI will personalise learning and improve efficiency, and more sceptical or risk-oriented perspectives that emphasise concerns such as overreliance on technology, inequitable access, or the opacity of algorithmic systems. With regard to school education, these competing narratives underline the need for more differentiated evidence on what AI-based tools actually *do* in educational settings, under what conditions they are effective, and how their effects vary between learners, subjects and the contexts in which they are implemented. It also creates a need

for evidence that can inform decisions in schools and education systems about whether, where and how such tools should be used in teaching and learning, as well as their implementation at school level.

In response to these needs, the report examines ITS and related AI-based adaptive learning tools in general education. Its primary focus is whether, and to what extent, these systems affect learning outcomes, motivation and engagement, as well as socio-emotional development. To interpret variations in these effects, the review also synthesises evidence on pedagogical mechanisms and implementation conditions. In this sense, the report adopts a focus that is deliberately narrower and more analytically tractable than that employed in broader discussions of AI in education. It responds to the need to understand more established forms of educational AI before drawing wider conclusions about AI in schooling more generally.

2.1. Defining intelligent tutoring systems

ITSs are generally understood as computer-based instructional systems that enable individualised and goal-directed interaction with learners. They do so by linking learner input to structured information embedded in the system in order to adapt outputs such as task selection, content sequencing and feedback. A defining feature of ITSs is the continuous adjustment of instruction through the use of learner data, in conjunction with pedagogical and domain-related structures represented within the system (Bernacki et al., 2021; Ma et al., 2014; Vogel, 2023).

ITS architectures typically comprise several interconnected core components (Holmes et al., 2019; Sottolare et al., 2014). A **learner model** represents information about the learner, such as current performance, progress, misconceptions or other characteristics relevant to adaptation (Holmes et al., 2019; Sottolare et al., 2014). A **domain or expert model** represents the subject content, target competences and problem structure on which instruction is based (Holmes et al., 2019; Sottolare et al., 2014). A **pedagogical or tutor model** represents the instructional strategies, feedback rules and support decisions through which the system responds to learner input (Holmes et al., 2019; Sottolare et al., 2014).

Additional components, such as interface, diagnostic or open learner models, may support interaction, the diagnosis of errors, or the transparency of decisions made by the system (Holmes et al., 2019; Ma et al., 2014). It should be acknowledged, however, that recent literature applies the term “ITS” with varying degrees of breadth. At times, it includes dialogue-based, AI-supported, hybrid or platform-based systems that provide adaptive and learner-responsive support but do not always make their underlying models explicit (Simbolon et al., 2025; Thomas et al., 2024b; Zhai et al., 2025).

To clarify the scope of this review, and in line with various studies (e.g. Mondia et al., 2023; Uriarte-Portillo et al., 2023; Wutthisarn & Junpeng, 2024), systems are therefore considered to be ITSs when they combine learner-responsive adaptation with instruction that is embedded in a specific subject domain. Systems that recommend content or generate feedback without a systematic connection to learner information and domain-specific instructional logic are not treated as ITSs.

It should also be noted that the terminology used in the empirical literature is not always consistent. Some studies examine systems that are functionally close to ITSs because they provide adaptive, learner-responsive support within a specific learning domain, but do not explicitly describe the architecture of these systems in terms of learner, domain and tutor models. For this reason, the present review also considers related, AI-based adaptive learning tools whose properties, as described, indicate functional equivalence to ITS, even if the canonical component terminology is not used.

For readability, the report uses ITS as the main abbreviation to describe all systems within this broader review scope. Unless otherwise specified, the term therefore covers both systems explicitly labelled as ITS, as well as related AI-based adaptive learning tools that meet the review's functional inclusion criteria. This usage reflects the broader review scope while preserving the conceptual distinction between canonical ITS architectures and functionally similar adaptive AI-supported learning systems (Holmes et al., 2019; Sottolare et al., 2014).

2.2. Conceptual impact framework

To guide the review, this report applies an **conceptual impact framework, developed by the authors**, that distinguishes between three analytically related dimensions: underlying mechanisms, moderating or contextual factors, and observed effects. This framework is intended as a heuristic orientation for structuring the review and for interpreting the empirical literature, not as a final explanatory model. Thus, the framework is intentionally broad at this stage of the report. The framework is later revisited and refined in Section 4.2, where the review findings are used to develop a more empirically informed synthesised impact framework.

The three-part structure of the framework reflects the review's *a priori* analytical rationale. Since the review is concerned not only with whether ITSs affect pupils' outcomes, but also with how they operate pedagogically and under what conditions their effects vary, the framework separates their underlying mechanisms, moderating or contextual factors, and observed effects. "Underlying mechanisms" captures the instructional and technological processes through which an ITS may influence learning. "Moderating and contextual factors" captures the learner, subject, setting and implementation conditions that may

shape how these mechanisms operate. “Observed effects” captures the elements within the dimension of observed effects, via which ITS-related changes are examined in the later synthesis.

The ordering of these dimensions is analytical rather than hierarchical. Underlying mechanisms are presented first because they describe how ITSs translate learner information into instructional support. Moderating and contextual factors follow because they shape how these mechanisms operate and whether they translate into outcomes. Observed effects are presented third because they represent the outcome domains in which the consequences of these processes are examined. The contextual dimension is therefore crucial because it avoids the observed effects from being attributed to the technology alone, and helps to explain variations between learners, subjects, settings and the conditions of implementation.

This distinction is also anchored in instructional design and ITS research. Instructional design research suggests that educational effects depend on how learning activities are structured, guided, supported through feedback, practised and integrated into meaningful learning processes, rather than on the presence of a particular medium alone (Merrill, 2002). This supports the focus on mechanisms. ITS research specifies relevant mechanisms in terms of learner modelling, domain representation, pedagogical decision-making, feedback, scaffolding, sequencing and learner–system interaction (Ma et al., 2014; Sottolare et al., 2014; VanLehn, 2011).

The dimension of observed effects is defined in parallel with the substantive focus of the review. Learning outcomes and effectiveness are included because ITS research has traditionally examined whether adaptive tutoring improves achievement, mastery, conceptual understanding, problem solving or skill development (Ma et al., 2014; VanLehn, 2011). Motivation and engagement form a further element of the dimension of observed effects because ITSs may also affect learners’ participation, persistence, interaction with feedback and support, willingness to practise, and continued use (Wang et al., 2025; Yu et al., 2023). The element well-being/socio-emotional outcomes is framed more cautiously, as newer ITSs and AI-supported learning environments increasingly address affective states, confidence, anxiety, frustration, cognitive load and emotional experience during learning (Hladký et al., 2025; Standen et al., 2020). At the same time, the expected effects of these mechanisms cannot be interpreted independently of implementation conditions, learner characteristics, subject domains, and the educational settings in which ITSs are used. The framework therefore treats mechanisms, context and outcomes as analytically distinct but interdependent dimensions.

The framework provides a structured lens for organising and interpreting the empirical literature on ITSs in general education. At its core, it assumes that ITSs do not affect the outcomes of pupils in an automatic or uniform way (Sottolare et al., 2014). Rather, their effects depend on how these systems shape learning

processes, as well as on the conditions under which they are implemented and used. Table 1 summarises this initial conceptual impact framework by setting out its three dimensions, their main elements, and their respective roles in the overall impact logic.

Table 1: Conceptual impact logic for ITSs

Dimension	Elements	Role in the impact logic
Underlying mechanisms	Feedback type and timing Adaptive sequencing Scaffolding strategies Technological design features	Mechanisms through which ITS shape learning processes by structuring instruction, guiding progression and responding to learner input
Moderating and contextual factors	Educational stage and age group Subject domain Educational context Teacher role and parental involvement	Conditions that influence the strength, direction, and consistency of effects
Observed effects	Learning outcomes / effectiveness Motivation and engagement Well-being / socio-emotional outcomes	Outcome domains in which the effects of ITSs become visible

Source: Authors' own analysis.

The first dimension deals with the underlying mechanisms via which ITSs may generate such effects. The underlying assumption is that outcomes are shaped not simply by the presence of a digital tool, but rather depend on how the system is embedded within the learning process, how it structures instruction, how it responds to input from learners, and how it guides progression through learning tasks. In this report, particular attention is given to feedback and timing, adaptive sequencing, scaffolding strategies and technological design features. These mechanisms are relevant because they represent core ways in which ITSs can influence learning processes – for example, by providing immediate feedback; step-by-step feedback; adjusting task difficulty or learning pathways; offering hints and step-by-step support; or shaping learners' interaction with the system through interface and design elements. In this sense, the framework treats ITSs not as a single type of intervention with uniform effects, but rather as a cluster of design and support mechanisms. These mechanisms influence outcomes in different ways over time, as learners engage with the system and as support is adjusted within specific instructional settings. This is consistent with design-oriented work on ITSs that conceptualises such systems in terms of instructional management and pedagogical response rather than as uniform technologies (Sottolare et al., 2014).

The second dimension captures moderating or contextual factors that may condition the extent, direction or consistency of these effects. Even where similar

mechanisms are used, their impact may vary depending on the learners and settings involved. The framework therefore includes educational stage and age group, subject domain and educational context, as well as the teacher's role and parental involvement. These factors are important because they shape both the implementation of ITSs and the capacity of pupils to benefit from them. For example, younger learners may respond to adaptive support differently from older learners; effects may differ between mathematics and language learning; use in the classroom may generate different dynamics from those in homework use; and the presence of teachers or parents may strengthen, mediate or constrain the ways in which ITSs are used. These contextual conditions also influence how mechanisms such as feedback, scaffolding or adaptive sequencing are enacted in practice. This emphasis on context is also in line with broader discussions on AI in education, which stress that its educational value depends not just on technical capability but also on the conditions under which it is implemented and the roles of educational actors (Holmes et al., 2019).

The third dimension contains the observed effects associated with the use of ITSs. In this review, these effects are grouped into three outcome domains: learning outcomes/effectiveness, motivation and engagement, and well-being or socio-emotional outcomes. Together, these reflect a broad understanding of educational impact. ITSs may influence not only pupils' academic performance and skill development, but also their willingness to engage with learning tasks, their persistence in the learning process and, where studied, broader affective or socio-emotional dimensions of learning. The framework therefore captures both cognitive and non-cognitive outcomes. Within this framework, the central place given to learning outcomes and effectiveness is strongly supported by meta-analytic evidence showing that ITSs are associated with measurable learning gains, even though the broader framework adopted here extends beyond achievement alone (Ma et al., 2014).

Overall, the framework assumes a dynamic relational logic: ITSs influence educational outcomes through specific mechanisms, while the strength and nature of these effects are conditioned by contextual and moderating factors. The framework understands ITS-supported learning as an iterative process in which learners, systems, teachers and (where relevant) parents, shape how support is enacted and how effects emerge over time. This perspective allows the review to move beyond the question of whether ITSs are, on average, effective, and to analyse more precisely how, for whom, and under which conditions their effects emerge. The framework thus provides the conceptual bridge between the review's descriptive mapping of the literature and its later synthesis of patterns in the evidence.

2.3. Purpose of this review

This review responds to a specific analytical and policy need, as it is intended to provide a comprehensive, evidence-based overview of the impact of ITSs in general education. More specifically, the report is designed to synthesise existing causal evidence on learning outcomes, motivation and engagement, as well as socio-emotional impacts, while also identifying underlying mechanisms and conditions for success.

A focused review of ITSs is particularly appropriate because the current evidence base on generative AI in education remains comparatively limited, especially with regard to causal evidence from school settings. By contrast, ITSs and related adaptive learning tools are more established forms of AI in education. Thus, they provide a stronger basis for examining how more traditional forms of AI may affect learning processes and educational outcomes in practice. This makes these technologies an especially suitable subject for an analytical report intended to inform evidence-based policy discussions.

The review is further justified by the fact that the available literature on ITSs is both substantial and heterogeneous. Studies differ in how they define ITSs. They also cover different subject domains and educational stages, examine different outcomes, and consider different mechanisms and contextual conditions. A systematic synthesis is therefore necessary in order to identify recurring patterns across this diverse evidence base, as well as to distinguish between more robust and less robust strands of evidence, and to clarify where conclusions remain conditional or uncertain.

Lastly, this review has a clear relevance for policy and practice. School systems are increasingly confronted with decisions as to where, whether, and how ITSs should be used (Hallström et al., 2026; Organisation for Economic Co-operation and Development [OECD] (2025). Such decisions require more than just isolated findings on average effectiveness. They require an understanding of which pedagogical mechanisms appear to matter most, under what implementation conditions positive effects are more likely to emerge, and how far evidence from non-European contexts can be interpreted for a European policy setting. The review therefore aims to provide a more differentiated evidence base for educational and policy decisions concerning the use of ITS in school education.

2.4. Research question

Against this background, the review is guided by the following research question:

What existing causal empirical evidence on intelligent tutoring systems (ITSs) and related AI-based educational technologies can be identified in general education

(primary to upper-secondary), regarding their impact on pupils' learning outcomes, motivation and engagement, and socio-emotional development?

This question is intentionally framed in a way that goes beyond average treatment effects. It is designed not only to identify whether positive, negative or mixed effects can be observed, but also to support a more differentiated interpretation of how such effects relate to pedagogical mechanisms and to the contextual conditions in which ITSs are implemented.

The relevance of the above question lies in the fact that debates on the use of ITSs in school education require more than isolated findings on overall effectiveness; they also require a clearer understanding of which outcome domains are supported by the available causal evidence, how consistent that evidence is, and where important limitations or uncertainties remain. The review therefore does not assume a single uniform pattern of impact, but instead seeks to determine whether the existing evidence points to differentiated effects across outcome domains, learner groups (educational stage and age) and implementation contexts.

2.5. Contribution

The main objective of this report is to provide a comprehensive and evidence-based synthesis of the impact of ITSs on pupil outcomes in general education. In doing so, the report contributes to a more differentiated understanding of AI in school education by focusing on a field in which a fairly substantial causal evidence base is already available.

More specifically, the report pursues four closely related objectives. First, it examines how the reviewed studies define and operationalise ITSs, including their core components, common features and different system types. Second, it identifies and synthesises the available evidence on learning outcomes and effectiveness, motivation and engagement, and – where evidence permits – well-being and socio-emotional outcomes. Third, it analyses the mechanisms and the moderating or contextual factors associated with the effects reported, paying particular attention to feedback, adaptive sequencing, scaffolding, technological design features, educational context, stage of education, subject domain, and the role of teachers and parents. Fourth, these findings are framed within a broader interpretive framework in order to clarify more precisely how, for whom, and under what conditions ITSs appear to contribute to pupil outcomes.

The contribution made by this review is therefore threefold. Empirically, it consolidates a diverse body of causal evidence on ITSs in school education. Conceptually, it links observed outcomes to pedagogical mechanisms and contextual conditions through the conceptual framework introduced in this chapter and the synthesised impact framework developed later in the report. Analytically, it supports a more precise interpretation of the educational potential

of ITSs by showing that their effects are neither automatic nor uniform, but depend on how adaptive instructional support is designed, implemented and embedded in practice. These conclusions also provide the basis for the report's final conclusions and implications.

3. Methods

To address the research question, this review follows a structured, systematic approach. This chapter describes the methodological design and procedures used to identify, screen, extract and synthesise the relevant empirical evidence.

3.1. Review design and protocol development

The literature review that informs this report followed a structured design, driven by a systematic literature review protocol. It encompassed the stages of search, screening, full-text eligibility assessment, data extraction, quality appraisal and synthesis. At the outset of the review, a research plan was developed to specify the methodological procedure in advance. This plan defined the overarching research question, established the screening criteria, and guided the development of the extraction matrix. The extraction matrix was constructed in alignment with a set of subordinate research questions, each addressing a particular descriptive or substantive dimension of the overarching review question.

3.2. Search strategy

The literature search was conducted using Elicit (<https://www.elicit.com>) as its primary scientific search tool. Elicit was selected because it is designed for structured academic literature searches, and operates on the basis of the Semantic Scholar database. Its principal function in the present review was to identify peer-reviewed empirical studies relevant to the review question. The search strategy was derived directly from the overarching research question, and focused on studies examining ITSs in school-related educational contexts. Elicit was used not only to identify potentially relevant studies, but also to support structured filtering on the basis of study characteristics and outcomes. In this sense, it served as the central platform for identifying evidence as well as early-stage assessment of its relevance.

3.3. Eligibility criteria

The strict eligibility criteria required that studies were published or otherwise made publicly available between 2020 and 2026, were written in English, and focused exclusively on learners in primary, lower-secondary or upper-secondary education. The starting point was set at 2020 to ensure that the review captured the most recent evidence, while ensuring the review's scope was feasible and

aligned with the rapid development of the field. Although the review focused on identifiable and methodologically transparent empirical studies, potential publication bias was addressed by complementing the database search with a targeted screening of working papers and preprints, as described in Section 2.9. In addition, studies were assessed against a set of criteria for their broader relevance. To be considered eligible, a study had to report at least one of the review's core outcome domains, namely learning outcomes or effectiveness, motivation and engagement, or well-being and socio-emotional outcomes. Furthermore, studies had to be empirical in nature and report data-based evaluation findings. The review prioritised intervention-oriented studies, including randomised controlled trials, quasi-experimental designs, controlled comparisons, pre/post intervention studies, and other structured evaluations of ITS use. Survey-based and self-report measures were deemed eligible only when they formed part of an otherwise eligible intervention-oriented study. Whether causal inference was possible depended on the design of a given intervention and its comparison structure, rather than on whether an outcome was measured through a test, questionnaire or self-report instrument. Studies based exclusively on survey, correlational or cross-sectional designs were excluded. Where survey-based or correlational analyses were reported within an included intervention study, these findings were treated as being descriptive or associational evidence and were not used to support causal conclusions regarding the effects of ITSs.

To be deemed eligible, studies also had to evaluate ITSs and be situated within a school-related educational context. This included the curricular use of ITSs in classroom instruction and/or homework, their extracurricular use within schools (e.g. as an additional learning support, for enrichment, or in after-school clubs), as well as non-institutional educational use such as self-directed or informal learning. A further condition for inclusion dealt with the operationalisation of ITSs. For the purposes of this review, an ITS was required to provide feedback and active support for student learning, rather than merely indicating whether answers were correct or incorrect. Studies were only included if core functions of the ITS model— namely, learner modelling, domain representation and pedagogical guidance – were identifiable in the described adaptive properties of the system, even where they were not explicitly named as separate components. Systems based on underlying generative AI models were excluded from the review.

3.4. Screening procedure

Screening was conducted in two stages: firstly of article abstracts, followed by full-text screening. In the first stage, Elicit was used to conduct an automated screening of 9,985 article abstracts. Each abstract was rated across all screening dimensions and assigned a relevance score between 0 and 5, with higher scores indicating greater relevance to the overarching research question. This automated scoring served as a mechanism for prioritising studies for manual review, in which the top 400 publications were manually assessed by two

reviewers in Elicit. At this point, the studies were evaluated against the predefined screening criteria as well as the operational definition of ITSs adopted for this review.

On the basis of this abstract-level assessment, 94 publications were considered sufficiently relevant to proceed to the second stage: full-text screening. Here, the full texts of these 94 studies were retrieved and screened manually by two reviewers using the same eligibility criteria applied previously at abstract level, now employed to assess the complete text of each publication. After full-text screening, 65 studies were deemed to meet the inclusion criteria and were retained for data extraction and quality appraisal.

3.5. Data extraction

Data extraction was carried out in Elicit using a predefined extraction matrix applied to all of the 65 studies selected. This matrix had been developed during the planning stage of the review and was explicitly linked to the subordinate research questions. Each extraction field therefore corresponded to a particular analytical dimension and captured either descriptive study information or content relevant to a specific aspect of the overarching research question. This procedure enabled evidence to be extracted in a systematic and comparable manner across the included studies. The completed extraction matrix formed the empirical basis for the subsequent synthesis.

3.6. Quality appraisal

Quality appraisal was conducted on all 65 of the included studies. Each study was assessed with regard to its internal validity, identification strategy, attrition risk, validity of measurement, and transparency. The purpose of this appraisal was to support a structured evaluation of the methodological robustness of the included evidence base and to inform the interpretation of the review findings during synthesis.

3.7. Synthesis

The synthesis of the extracted evidence was supported by the GPT-5.4 model. More specifically, the individual columns of the extraction matrix were used to address distinct sub-dimensions of the overarching research question. On this basis, the extracted content was processed through iterative summarisation, clustering and transformation into academic prose. These synthesised outputs served as an analytical foundation for the reviewers, who used them to develop the individual chapters of the publication. The synthesis process thus combined

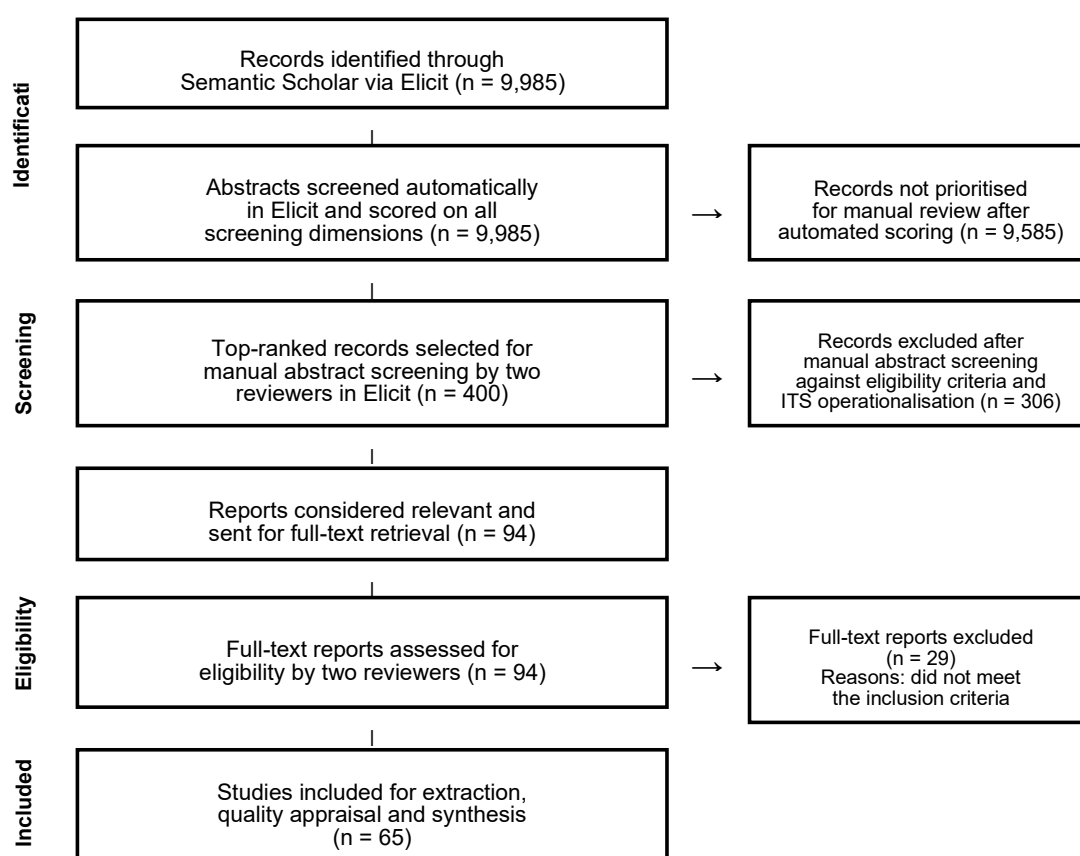
structured AI-assisted evidence processing with reviewer-led interpretation and scholarly writing.

3.8. Adapted PRISMA flow diagram

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) provides a structured reporting framework that ensures transparency and reproducibility in documenting the identification, screening, eligibility assessment and inclusion of studies within a systematic review. Figure 1 presents the study selection process using an adapted PRISMA flow diagram.

A total of 9,985 records entered the screening process. Following automated prioritisation, 400 records were selected for manual screening of abstracts, with 306 of these being excluded on the basis of the predefined criteria. This resulted in 94 studies proceeding to full-text assessment. All 94 full-text articles were assessed for eligibility, leading to the exclusion of 29 studies. The remaining 65 studies met all of the inclusion criteria and were included in the final sample for data extraction, quality appraisal and synthesis.

Figure 1: Adapted PRISMA flow diagram



Source: Authors' own analysis.

3.9. Working papers

To complement the database search with recent unpublished or pre-publication evidence, the review also screened working papers and preprints from three relevant sources: NBER, OSF/EdArXiv and SSRN. This step was intended to capture emerging empirical studies that may not yet be available as journal publications, while maintaining the review's focus on methodological transparency and substantive relevance. In line with the review protocol, working papers were only considered where they were clearly identifiable and methodologically transparent.

As a result, only one working paper was retained for further consideration after validation. This paper, Borchers et al. (2023), compared ITS-supported and paper-based mathematics practice in school settings and found that the ITS provided more practice opportunities that pupils eventually answered correctly, but did not produce stronger overall learning gains, with effects varying across curricular units.

The search of working papers primarily served to check whether the review had adequately captured the most recent developments in the field. Given that only one eligible working paper was retained (and this was published in 2023), this supplementary search did not add substantively new unpublished evidence beyond the reviewed literature, but it did help to confirm that no more recent working papers meeting the review's definition of ITS were identified.

4. Results

This chapter of the report presents the results of the systematic review, including the characteristics of the studies included, the definition and operationalisation of intelligent tutoring systems (ITSs), and the main empirical findings across outcome domains, mechanisms and contextual factors.

4.1. Descriptive overview of the reviewed studies

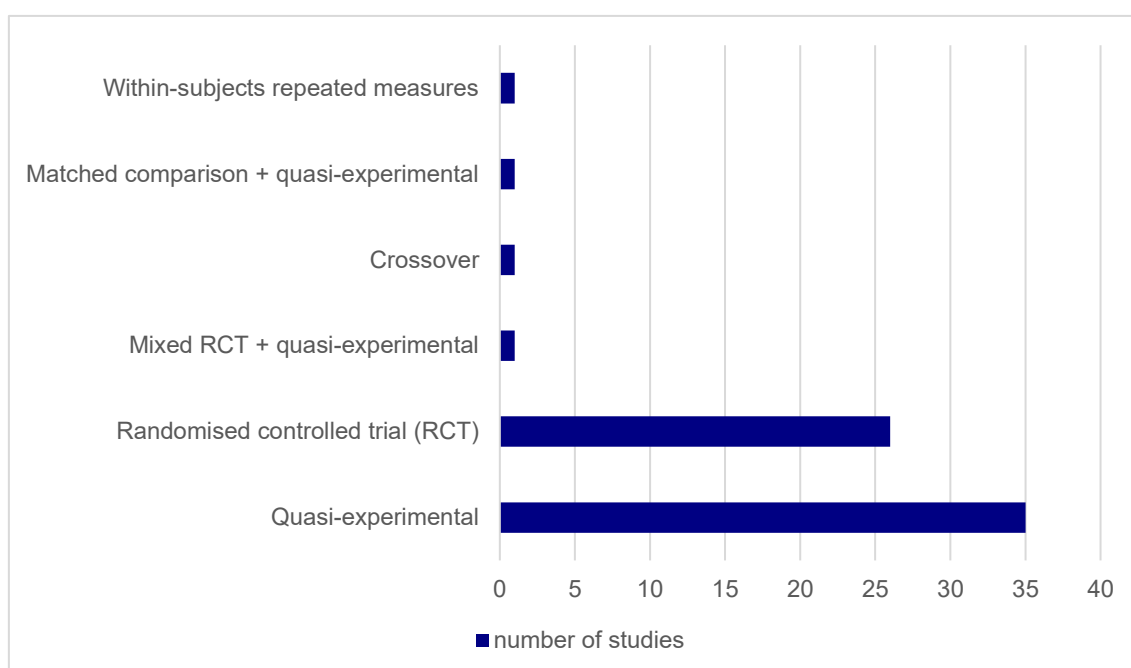
This section presents the main characteristics of the studies reviewed in terms of study design, countries and regions, time of research (publication and data collection), school level and age group as well as examined outcomes.

4.1.1. Study design

The evidence base can be described as predominantly intervention-oriented, with most studies aiming to evaluate the effects of ITSs on learner outcomes. As shown in Figure 2, the largest group of studies used quasi-experimental designs, followed by randomised controlled trials. Together, these two categories of design account for the majority of the studies included. The remaining design types occur much less frequently and can be interpreted as smaller methodological subgroups within the review evidence. These include cross-over designs, matched comparisons and quasi experimental designs, mixed RCT/quasi-experimental designs, and within-subject repeated-measures designs. Within the quasi-experimental category, various methodological approaches were used. Most studies employed non-randomised comparison structures, comparing an intervention group with a control or comparison group, an alternative instructional condition, or another version of the tutoring system.

In 27 of the 35 studies coded as quasi-experimental, a control or comparison group was explicitly reported. These comparison conditions included regular instruction, traditional classroom teaching, business-as-usual instruction, alternative digital learning tools, non-ITS environments, or comparable learner groups. Other studies used related comparison approaches, including matched groups, naturally existing classes, within-subject pre/post comparisons, repeated measurements, or comparisons between different system conditions. Overall, the category of quasi-experimental designs therefore comprises a range of non-randomised intervention evaluations with different comparison structures.

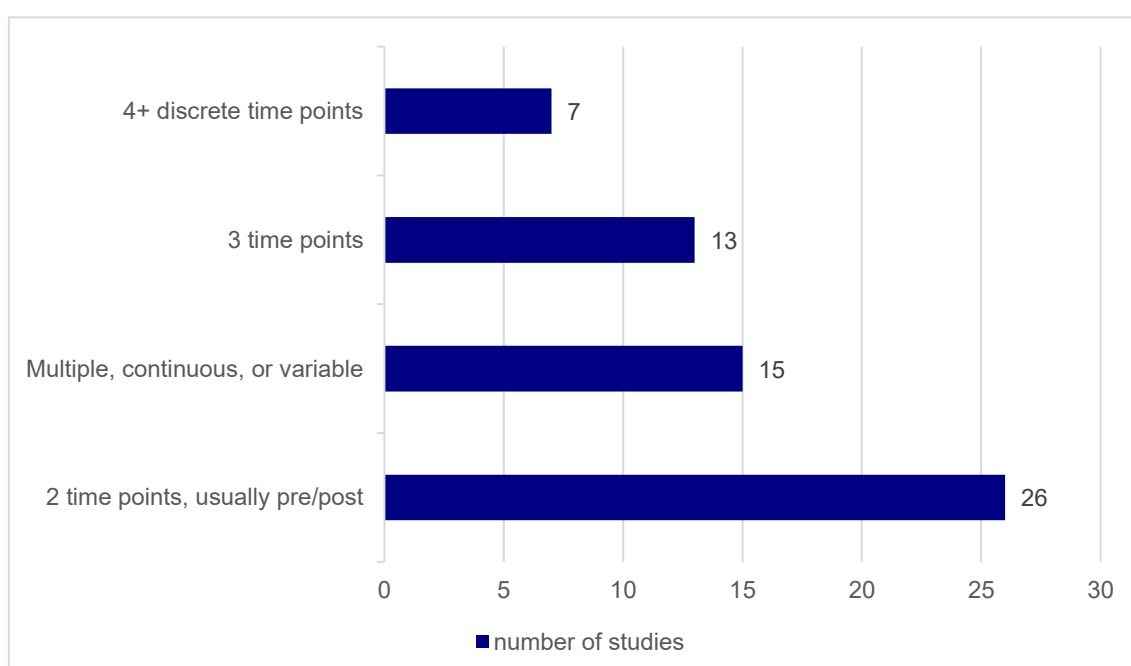
Figure 2: Distribution of study design types across included studies



Source: Authors' own analysis.

The predominant structure for evaluation was **comparative and intervention-**focused, often relying on pre/post or baseline/endline measurements. As shown in Figure 3, the most common measurement structure involved two time points, typically before and after the intervention. Other studies used three or more discrete measurement points, while some relied on multiple, continuous or variable data collection points. For four studies, the publications did not provide sufficient information to classify the number or timing of measurement points.

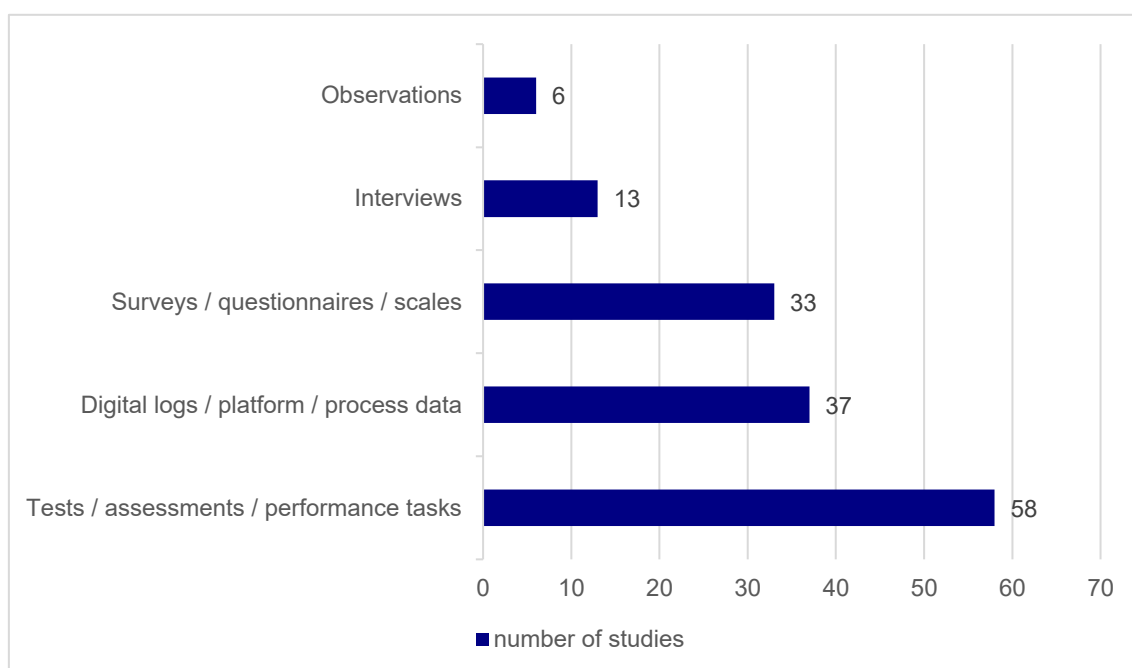
Figure 3: Distribution of measurement point used



Source: Authors' own analysis.

The most frequently used research instruments in the studies reviewed were **tests, assessments or performance tasks**. These were often complemented by digital logs, platform data, process data, surveys, questionnaires, scales, interviews and observations. Figure 4 shows that tests and performance-based measures were the family of measurement tools most commonly used, followed by digital trace data and survey-based instruments. Most studies employed more than one research instrument; thus, they are counted more than once in Figure 4.

Figure 4: Distribution of research instrument families



Source: Authors' own analysis.

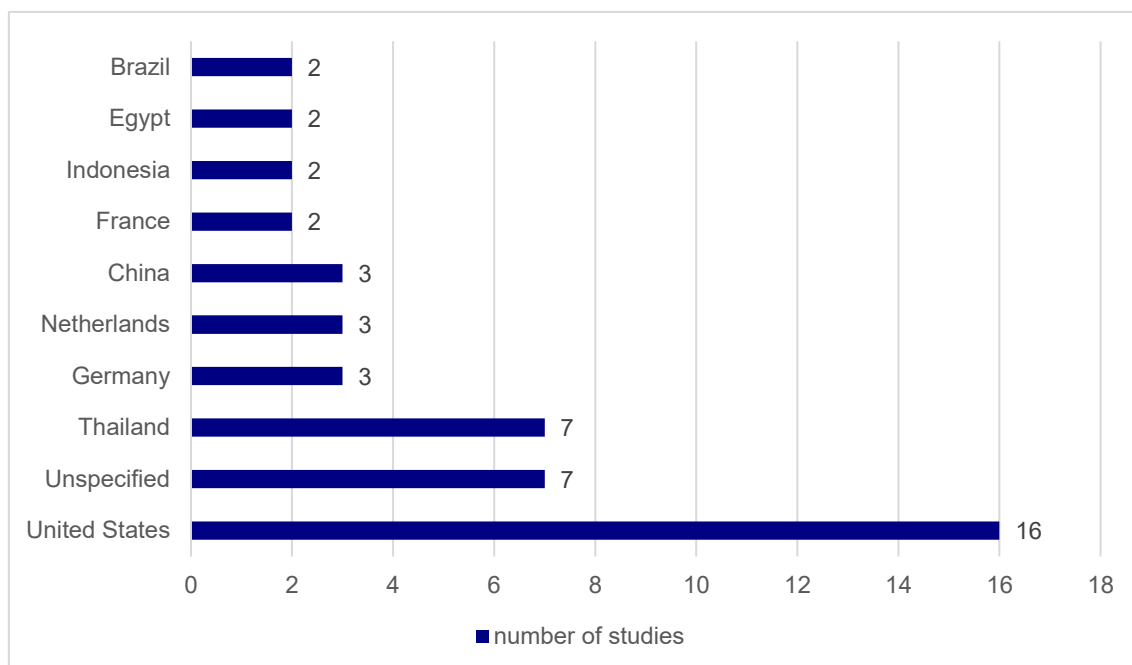
In summary, the evidence reviewed is mainly based on intervention-oriented studies – in particular, quasi-experimental designs and randomised controlled trials. Most studies used comparative or pre/post evaluation structures, with the most common instruments being tests or performance tasks. These were supplemented by digital trace data and surveys. The methodological strength of the evidence varies between studies depending on features such as randomisation, the presence and type of control or comparison condition, matching procedures, repeated measurement, and the nature of the outcome instruments.

4.1.2. Countries where research was conducted

The country in which the largest number of studies were conducted is the United States, which accounts for a quarter of the studies, followed by Thailand, Germany, China and the Netherlands (see Figure 5). At a regional level, however,

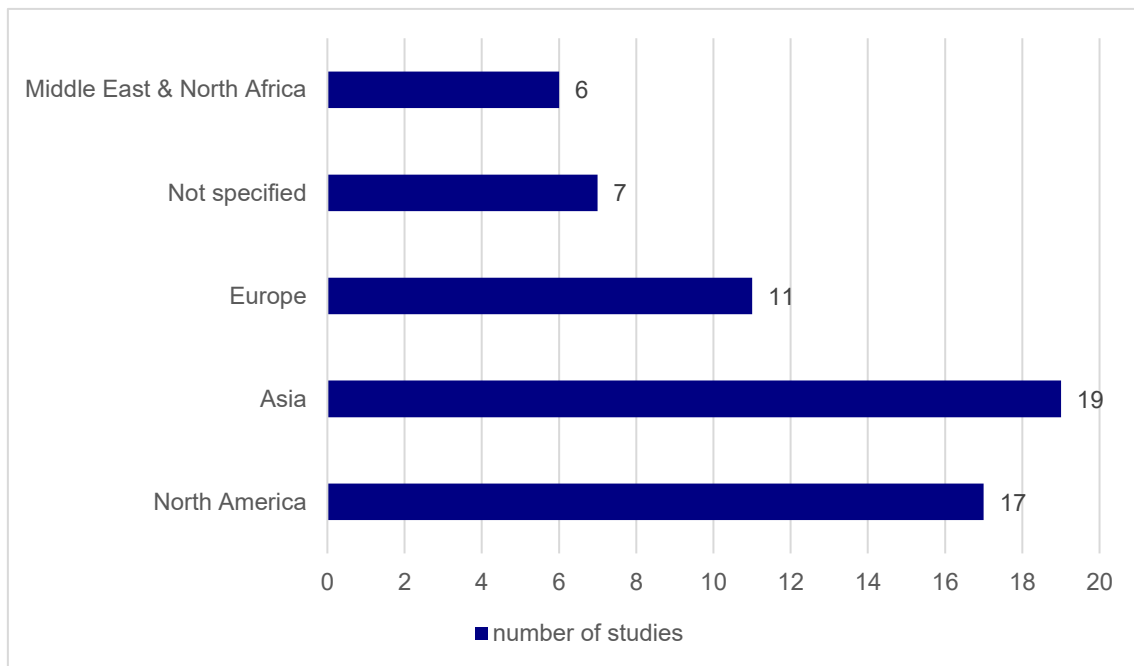
Asia is the continent in which the majority of the included studies were conducted (see Figure 6).

Figure 5: Top countries represented in the evidence base



Source: Authors' own analysis.

The field therefore focuses on North America – in particular, the United States – followed by Southeast Asia (in which Thailand stands out as the most prominent contributing country), and Europe (represented by Germany, the Netherlands and France). Evidence from Latin America, South Asia, Central Asia, Sub-Saharan Africa, and Russia/Europe-Central Asia is sparse, usually represented by only one or two studies. Thus, the evidence appears skewed towards only a few countries. Countries that are only represented once are not shown in Figure 5.

Figure 6: Regional distribution in the evidence base

Source: Authors' own analysis.

As the figure shows, the geographical distribution of the evidence base is uneven, with the United States and selected Asian countries being particularly strongly represented. The implications of this imbalance for interpreting the transferability of findings to European school systems are discussed in Section 5.3, “Transferability caveats for non-EU studies”.

4.1.3. Timing of research

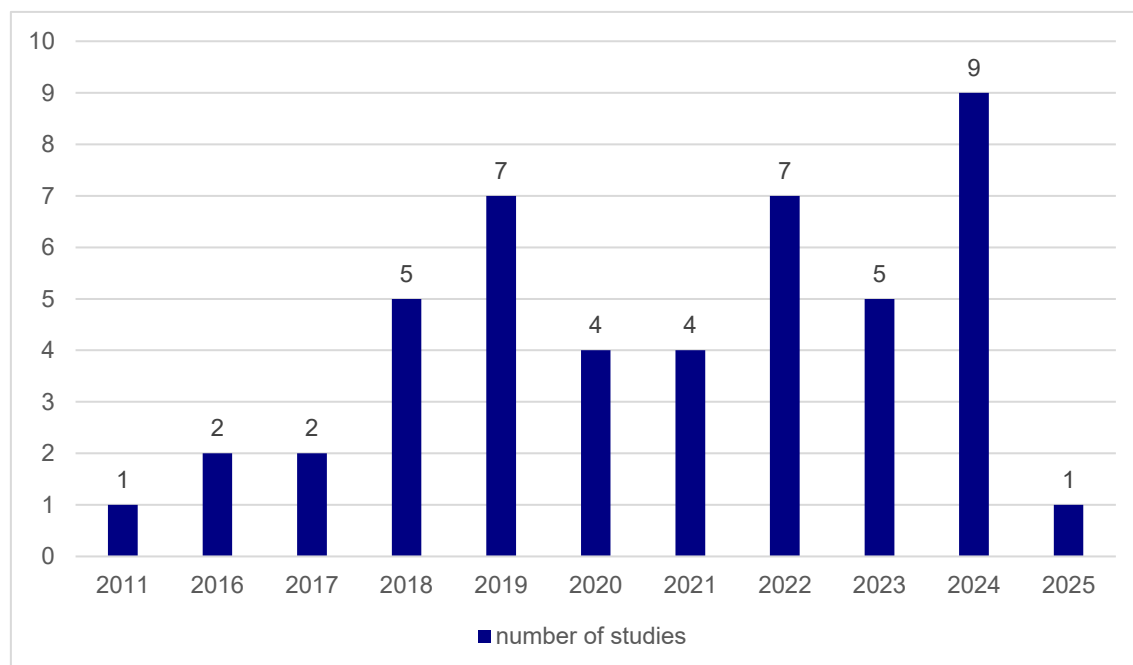
Among the studies included, publishing activity has increased significantly over time. The dataset begins with six publications in 2020, dips to four in 2021, then rises steadily: five in 2022, eight in 2023, 16 in 2024, and 25 in 2025. This means that the years 2024 and 2025 alone account for 62.1 % of all records, indicating a strong recent acceleration in publishing activity. The value for 2026 is only two publications, due to the year being only partially complete at the time of review.

However, when one considers the time at which the studies were actually carried out or the data collected, the above trend appears less striking, as 37 out of the 65 publications (56.9 %) did not identify a specific year in which data were collected. Among the 28 studies that did include identifiable dates, activity was sparse prior to 2016, increased around 2018–2019, remained constant between 2020 and 2023, and reached its highest observed level in 2024. In the case of long-term studies, each year covered by the study was counted (see Figure 7).

The key interpretation of this is that the field appears to have moved from sporadic earlier studies to a more active and rapidly expanding research area, particularly after 2020. However, the large share of studies with missing or unclear data

collection dates means that this trend should be read as a lower-bound estimate, not as a complete historical record.

Figure 7: Research activity by year of data collection

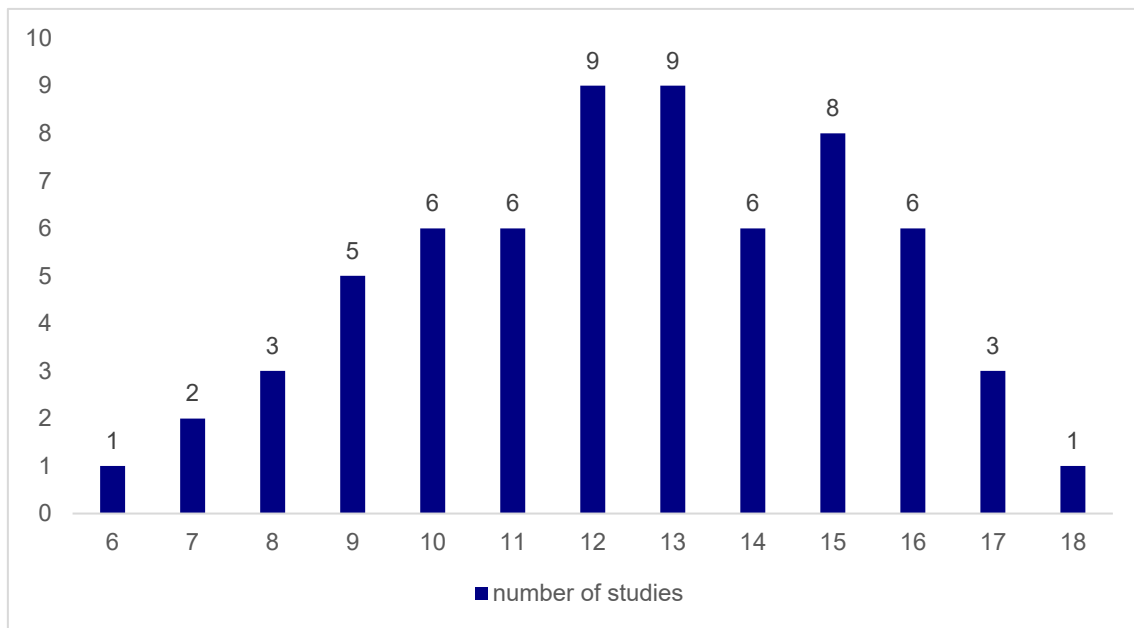


Source: Authors' own analysis.

4.1.4. School level and age group

More than half of the studies included (35; 53.8 %) focus on the use of ITSs in lower-secondary level, followed by primary schools (21; 32.3 %) and upper-secondary level (17; 26.1 %). Among these three groups, eight studies span multiple educational levels

Information on the age of the pupils studied is available for only 22 studies. Most pupils were in early adolescence, in particular around age 12–15. This aligns with most studies being conducted at lower-secondary level (see Figure 8). Note that in many cases, one study can contribute to several year or age bands.

Figure 8: Age coverage in studies providing explicit age information

Source: Authors' own analysis.

4.1.5. Subject areas covered

Most studies examined the effects using ITSs in mathematics, followed by language/reading and sciences. Because some studies focused on more than one subject, the total number of subjects covered, as reported in Table 2, exceeds the total number of included studies (65).

Table 2: Distribution of subject areas across included studies

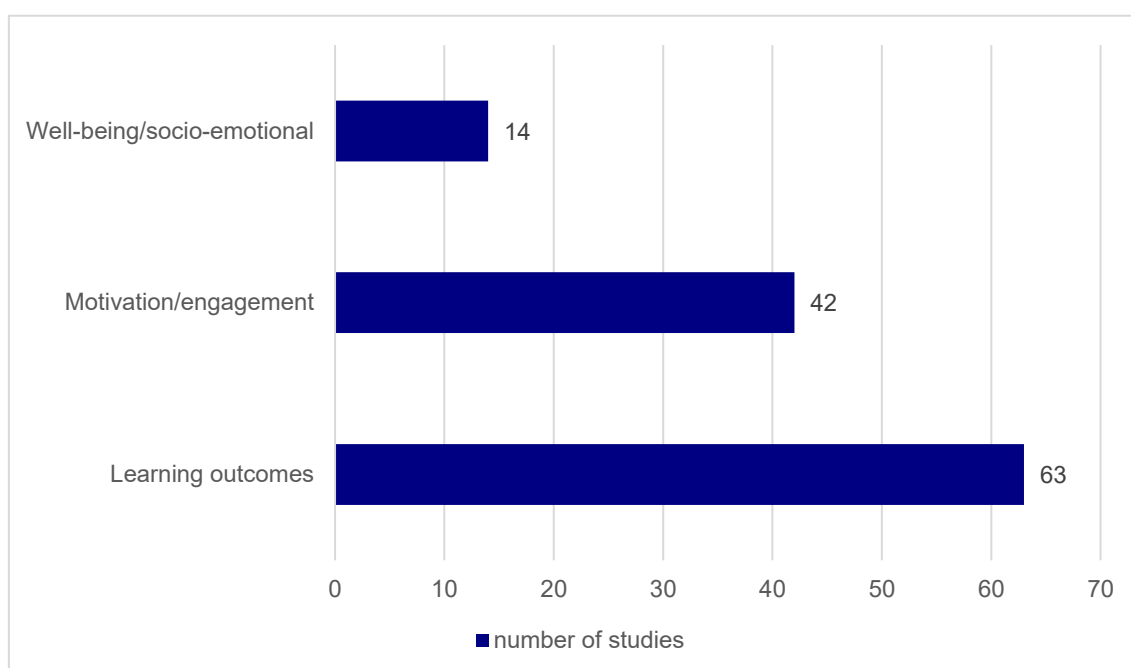
Subject group	Reports in which it is mentioned
Mathematics	41
Language / reading / literacy	16
Science	12
Computer science	4
Physics	3
Biology	2
General / cross-curricular	2
Chemistry	1
Social studies	1
Sports	1

Source: Authors' own analysis.

4.1.6. Outcomes examined

All studies except for two (97 %) examined the effects of ITSs on learning outcomes. In addition, around 63.6 % examined motivation/engagement, and 21.2 % examined well-being/socio-emotional outcomes. The exact distribution of these focus areas can be seen in Figure 9. Because the above dimensions are not mutually exclusive, one publication may count towards multiple categories. The most common pattern is learning outcomes + motivation/engagement, which appears in 32 publications. This is followed by learning outcomes only (which appeared in 19 publications). All three dimensions are examined in only eight publications (12.1 % of the total).

Figure 9: Distribution of outcome dimensions examined



Source: Authors' own analysis.

4.2. Definition and types of ITSs

This section of the report examines how ITSs are defined and operationalised in the studies reviewed. It identifies their core components, recurring features and distinguishable system types.

4.2.1. Operationalisation of ITSs in the studies reviewed

Among the studies included in this review, ITSs are most commonly defined as computer-based or AI-based instructional systems that personalise learning by adapting content, support, pace or difficulty to the learner, and by providing

individualised feedback (Khazanchi et al., 2024; Uriarte-Portillo et al., 2023). A closely related formulation describes ITSs as systems designed to simulate key functions of human tutoring – in particular, one-on-one support – by offering tailored guidance and adaptive feedback (Olney et al., 2025; Pervaiz et al., 2025). Overall, the extracted findings indicate a shared conceptual core that centres on adaptivity, personalisation and support that is responsive to the learner. Most of the studies operationalise ITSs in terms model-based architectures that combine components such as a domain model, learner model, tutoring model and interface model (Alakayleh, 2025; Mondia et al., 2023).

Another recurrent theme is the operationalisation of ITSs in terms of immediate instructional support – especially feedback, hints, scaffolding, guided practice and step-level assistance (Azevedo et al., 2020; Borchers et al., 2023). Several studies define ITSs more specifically in terms of dialogue-based tutoring, including conversational agents, chatbots or natural language tutors that guide learners through questioning, explanation, or error-based support (Liu et al., 2025; Shih et al., 2023). In these cases, ITSs are framed as an interactive tutoring process rather than just a mechanism for content sequencing. Thus, ITSs are frequently defined not only as an instructional tool, but also as a system that continuously adjusts instruction on the basis of learner data and estimated learner state. In this context, “learner state” refers to the system’s inferred representation of a learner’s current situation. This includes aspects such as knowledge, performance, progress, errors, needs or engagement, which can be used to inform adaptive tutoring decisions (Sottolare et al., 2014).

Despite the shared core of adaptive features, the scope of the concept of ITS varies between studies. Some publications apply the term to relatively conventional adaptive tutoring software (Baker et al., 2020; Kim et al., 2020), whereas others extend it to hybrid human–AI tutoring, mixed-reality systems, robotics, affect-sensitive systems, gamified systems or collaborative environments (Standen et al., 2020; Thomas et al., 2024b). In addition, some studies treat established platforms as ITSs or ITS-like systems, while others discuss adaptive learning technologies more broadly without clearly delimiting the concept. A small number of studies do not explicitly define ITSs directly – a further indication of variations in conceptual precision within the field (Majumdar et al., 2023; Oreopoulos et al., 2024).

Taken together, the evidence shows that the literature is consistent in defining ITSs at a general level as being adaptive and personalised systems that provide responsive instructional support based on learner characteristics or performance (Mondia et al., 2023; Khazanchi et al., 2024). However, the studies diverge in how narrowly or broadly they delimit the concept, including whether ITSs are treated as a model-based tutoring architecture, a dialogue system, a specific platform, a hybrid human-AI arrangement, or a broader category of adaptive learning technology (Mondia et al., 2023; Thomas et al., 2024b). The evidence therefore supports a common conceptual core, but not a fully uniform definition that applies across the field. ITSs are therefore generally understood as adaptive,

individualised, data-informed tutoring systems, while their operationalisation varies across technologies, domains, architectures and levels of automation.

4.2.2. Core components of ITSs in the studies reviewed

ITSs are commonly described as comprising four core components: an expert or domain model, a learner model, a tutor or pedagogical model, and a communication model. In this review, all 65 studies covered these components at least implicitly through their descriptions of domain knowledge, learner-related data use, instructional decision-making and learner–system interaction. However, only 33 studies explicitly report all four core components as distinct elements of the ITS architecture. A further 20 studies explicitly report the learner, tutor and communication models, while covering the expert or domain model functionally but not articulating it as a separate component (Mondia et al., 2023; Morais & Jaques, 2022; Shih et al., 2021; Uriarte-Portillo et al., 2023; Wang et al., 2024; Wutthisarn & Junpeng, 2024).

Where **expert components** were explicitly described, they were most often framed as domain representations, knowledge structures, curriculum-aligned content models or knowledge bases ranging from general content models to more formal knowledge-structure and inference-based models (Borchers et al., 2023; Uriarte-Portillo et al., 2023; Wutthisarn & Junpeng, 2024). As noted above, however, this component was less consistently foregrounded than the learner, tutor, and communication components.

The **learner model** was one of the most consistently represented components and was commonly operationalised in terms of profiling, diagnosis, progress monitoring, and personalisation. Among the studies reviewed, learner components were described as tracking progress, difficulties, performance, proficiency tiers, or individualised pathways and profiles (Khazanchi et al., 2024; Marwiani et al., 2025; Thomas et al., 2024b). Even where a learner model was not directly labelled as such, studies often connected learner modelling to adaptive guidance or logged records of learners' actions and responses within the system reinforcing the centrality of learner-state representation in the included literature. More formal learner modelling is also reported through diagnostic decision trees (Verma & Kulkarni, 2025) and Bayesian knowledge tracing (Huang et al., 2021).

The **tutor model** was the most consistently explicit in the dataset, and was typically described as the mechanism for instructional control, scaffolding, feedback and sequencing. Many records referred to problem assignment, hint generation, automated or real-time feedback, modelling, practice, assessment, scaffolding and next-step guidance (Bhatt et al., 2024; Borchers et al., 2023; Wijekumar et al., 2024). Some studies also described the functions of tutor models in broader terms, including support in planning and orchestration, motivational support or agent-based guidance (such as the use of pedagogical

agents). Overall, the tutor model is the most consistently apparent and operationally visible ITS component (Feng et al., 2025; Hladký et al., 2025; Oreopoulos et al., 2024; Thomas et al., 2024b).

The **communication model** also occurred frequently, and was represented by dashboards, interfaces, feedback channels and dialogue systems. A substantial group of studies described dashboards, teacher-facing analytics, progress displays and other interface elements that made learner status and instructional responses visible to tutors, teachers or the pupils themselves (Khazanchi et al., 2024; Oreopoulos et al., 2024; Thomas et al., 2024b). Another recurring pattern involved dialogic or conversational interfaces, dialogue-based interaction and multilingual conversational formats (Liu et al., 2025; Olney et al., 2025). In addition, some studies described multimodal, visualisation-based or augmented reality-oriented communication functions (Abdulayeva et al., 2025; Uriarte-Portillo et al., 2023; Zhai et al., 2025).

These findings indicate that the communication component was commonly tied to making system decisions observable and interactive rather than being limited to a narrow user interface layer. In this sense, communication models translate system decisions and learner data into visible forms of feedback, monitoring and interaction that teachers and learners can understand and act upon.

Where studies clearly specified the **expert component**, the architecture more closely resembled the canonical four-component ITS model (e.g. Mondia et al., 2023; Uriarte-Portillo et al., 2023; Wutthisarn & Junpeng, 2024). When the expert component was absent or only implicit, the remaining triad of learner, tutor and communication components still appeared with great regularity and carried most of the functional description of adaptation, support and interaction (Thampitak et al., 2025; Verma & Kulkarni, 2025; Zhai et al., 2025).

The canonical four-component architecture therefore remains common, but is not universal. Where the expert model is not implicit, the resulting learner–tutor–communication models still carry most of the functional description of adaptation, support, and interaction.

4.2.3. Features of ITSs in the studies reviewed

This section examines the features of ITSs that are most frequently reported in the studies reviewed, such as adaptive learning, immediate and specific feedback, and sequencing. It goes on to analyse how these features were used to adapt instruction, guide learner performance, monitor progress, support engagement and facilitate implementation.

Across the publications, the most frequently named features in studies were adaptive learning (63.2 %), sequencing (52.9 %), and immediate and specific feedback (51.5 %). When closely related labels are grouped together, adaptive

or personalised provision and feedback or hints appeared in 97.1 % of studies; sequencing, mastery or pacing in 76.5 %; assessment, monitoring or analytics in 72.1 %; interactive, multimedia or conversational interaction in 50.0 %; socio-emotional, motivational or metacognitive support in 33.8 %; teacher-facing support in 27.9 %; and explicit scaffolding in 22.1 %. The studies reviewed therefore focus mainly on instructional adaptation, feedback and the management of progression as the main features of ITSs.

The ITSs studied were most frequently described as the adaptation of content, difficulty, activities, pace or learning pathways to learner performance, prior knowledge, understanding, readiness, grade level, strengths and weaknesses, effort, affect or learner profiles (Simbolon et al., 2025; Thomas et al., 2024b). In several studies, this adaptation was operationalised through formal mechanisms such as Bayesian knowledge tracing (Yannier et al., 2024), fuzzy logic, decision trees, supervised machine learning, reinforcement learning and multi-armed bandit algorithms (Clément et al., 2025; Uriarte-Portillo et al., 2023). Across the studies, personalisation also extended beyond task difficulty to the sequencing of learning routes, the selection of educational resources and practice exercises, content delivery and the timing or form of instructional support, with some systems incorporating multimodal inputs, prior Moodle activity, demographic information, self-esteem, emotional quotient or socioeconomic stratification in the tailoring process (Ezzaim et al., 2023, 2024). Some systems used more formal modelling approaches or additional learner-profile information, but these details are not reported consistently and are therefore treated here as examples of broader adaptive provision rather than as separate categories of features.

Feedback was typically described as immediate, specific, real-time, automated, or personalised, and was delivered via hints, explanations, examples or step-level validation (Morais & Jaques, 2022; Pervaiz et al., 2025). Several studies explicitly linked feedback to the diagnosis of misconceptions, error patterns, invalid actions or guessing behaviour, and some systems used feedback loops to recommend videos, practice tasks or more suitable learning material (Bringula et al., 2025; Huang et al., 2020). Scaffolding appeared as a related but more specific cluster. This ranged from prompts that encouraged reasoning to adaptive scaffolding, scaffolded dialogue, scaffolded handwriting support and hints available on demand with graduated specificity (Morais & Jaques, 2022; Munshi et al., 2022; Su et al., 2025). Taken together, feedback and scaffolding were the main features through which systems provided learner-responsive support during task completion.

Sequencing functions regulated how learners progressed through the material. The systems selected the next problem, activity, lesson, text or hint according to the student's mastery, readiness, progress, response time, error patterns, effort or diagnostic information. Some systems incorporated repeated practice, repetition, mastery thresholds or controlled exploration before learners were allowed to advance (Pervaiz et al., 2025; Huang, 2025; Yannier et al., 2024). Sequencing could be strongly structured, as in staged workflows, increasing

levels of difficulty, curriculum-aligned modules, equation plans or texts ordered according to the required level of mastery, but the studies reviewed also included flexible forms such as freely ordered modules or multiple valid solution paths (Azevedo et al., 2020; Su et al., 2025; Wijekumar et al., 2019). This indicates that sequencing was used both to structure progression through predefined learning materials and to adapt learning pathways to pupils' current performance or needs.

Assessment, monitoring and analytics formed an important informational layer across the corpus. The studies described diagnostic assessments, dashboards, progress checks, learning analytics, learner profiles, learner models, detailed performance metrics, interaction logs, real-time tracking and real-time data collection (Katchapakirin et al., 2022; Khazanchi et al., 2024; Huang, 2025). These monitoring functions were frequently linked to management by teachers or tutors through dashboards, reports, progress screens, manual controls and diagnostic reporting that could inform pacing, classroom instruction, or intervention decisions (Khazanchi et al., 2024; Thomas et al., 2024b). In a smaller number of studies, monitoring also drew on multimodal indicators such as speech, image, video, handwriting or facial expression, but these features were less common than conventional performance and progress data.

A further cluster of reported features relate to **interface design and interaction mode**. Many ITSs used multimedia, chatbots, conversational agents, speech technologies, image-based interaction, gamified environments, guiding characters, kid-friendly interfaces or visual aids to organise and deliver support (Katchapakirin et al., 2022; Liu et al., 2025). A smaller but recurring set of functions addressed socio-emotional, motivational and metacognitive support. This included emotional support, empathetic or motivational messaging, learning companions, goal-setting, self-explanation, self-evaluation, mindfulness exercises, reduced embarrassment and progress visualisation for metacognition (Feng et al., 2025; Hladký et al., 2025; Wang et al., 2025). The material extracted in this column predominantly described system functions and implementation processes; explicit empirical effects were not a major focus of the recorded feature statements.

Overall, the reported features served five main instructional purposes. First, adaptation and sequencing were used to **match instruction to learner state** by adjusting content, difficulty, pace, pathways and practice opportunities. Second, feedback and scaffolding were used **to guide performance during task completion** by correcting errors, identifying misconceptions, validating steps and providing next-step support. Third, **assessment, monitoring and analytics** were used to make learner progress visible to the system, teachers or tutors, and to inform instructional decisions. Fourth, **interface and interaction** features were used to make learning activities more accessible, usable or engaging. Fifth, **teacher-facing** tools were used to support implementation and classroom orchestration. These purposes are reported primarily as operational functions of the systems; their effects on learning, motivation, and socio-emotional outcomes are examined in the sections that follow.

4.2.4. Types of ITSs in the studies reviewed

The findings identify several distinct types of ITSs. These can be distinguished on the basis of the recurring dimensions, including interaction format, adaptation and learner-modelling logic, computational method, delivery modality, and the focus of support, including cognitive, affective, socio-emotional and gamified support (Thomas et al., 2024b).

The first distinction focuses on the **instructional domain**. The dataset includes web-based and text-focused systems such as the ITS for the “text structure strategy”, which is described as individualised-adaptive and is designed to help pupils identify text structures in order to produce their main ideas and summaries; one source also reports the use of an animated pedagogical agent (Wijekumar et al., 2019; Wijekumar et al., 2024). Language and literacy variants include an automatic speech recognition-based reading tutor for Dutch first-graders (Bai et al., 2022), an AI-powered foreign language learning app (Yan, 2024), an AI conversational agent from LAIX that provides personalised feedback, speech evaluation and learning assessment (Wang et al., 2025), and a dialogic tutor named SingaKids (Liu et al., 2025). In the US context, mathematics across grades 6–12, including first-degree equations, measurement, and geometry, as well as high-school biology, physics, and programming (Marwiang et al., 2025; Olney et al., 2025; Thomas et al., 2024b).

A further type is defined by **conversational interaction**. Categories here include one-on-one, dialogue-based mathematics tutors (Shih, Chang et al., 2023); a dialogue-based biology tutor named Guru, developed on the model of human subject teachers (Olney et al., 2025); the ScratchThAI framework, which involved a virtual conversational chatbot and automated assessment (Katchapakirin et al., 2022); and a dialogic support system for mathematics based on the AutoTutor ITS (Shih et al., 2023).

Another major cluster consists of adaptive and diagnostic ITSs that are defined by the learner data they process and the instructional adjustments they make. A shared methodological feature is the use of data-driven approaches to **diagnose learner needs** and to personalise instruction. One system integrates machine learning based on cognitive load theory to create a three-tier, adaptive cognitive load regulation framework that uses reinforcement learning and ensemble learning for real-time cognitive state detection, personalised intervention and metacognitive awareness (Zhai et al., 2025). Another combines decision trees, reinforcement learning and learner analytics (Verma & Kulkarni, 2025), while another “adaptive learning companion” uses natural language processing and reinforcement learning algorithms (Anafa & Friday, 2025; Verma & Kulkarni, 2025). Diagnostic systems based on machine learning or decision trees classify learners, estimate competency or proficiency, and use these to generate customised pathways, supplementary lessons or recommendations. This group includes Junpeng’s machine learning-based diagnostic system (Prasertsang et al., 2025a); a Python decision-tree model for scientific competency diagnosis

(Wutthisarn & Junpeng, 2024); a digital environment for measurement and geometry that uses supervised decision-tree classification; and an intelligent personalised diagnostic and tutorial system delivered via a digital platform (Prasertsang et al., 2025b). Taken together, these examples show that adaptive diagnostic ITSs use learner data not simply to assess performance, but to convert diagnostic inferences into personalised instructional actions such as feedback, pathway adaptation and targeted recommendations.

Within the dataset, a broader set of AI-driven adaptive systems can be identified that **personalise content, sequencing or feedback** on the basis of progress data or learner profiles. These include an interactive digital module that integrates socio-cultural content and delivers immediate feedback (Simbolon et al., 2025); a personalised study guide Moodle plug-in identified as an AI-based adaptive learning system; and an adaptive platform that tracks response times and errors to generate personalised learning paths (Ezzaim et al., 2024). Other systems used self-esteem, emotional intelligence and demographic variables to predict performance and adjust learning while incorporating methods such as “Teaching at the Right Level” and the flipped classroom approach, or provide adaptive sequencing, formative feedback, mastery thresholds, hints on demand, spaced review and intelligent diagnostic feedback (Ezzaim et al., 2023; Pervaiz et al., 2025).

Another group is defined by **explicit computational or representational methods**. Some systems are described through Bayesian knowledge tracing (Yannier et al., 2024), fuzzy logic (Uriarte-Portillo et al., 2023), Performance Factors Analysis, Markov decision process modelling (Mondia et al., 2023), fuzzy Bayesian modelling (Wen & Pan, 2025) or hypergraph representation (Olmo-Muñoz et al., 2022). In some cases, these methodological orientations are paired with distinctive modes of delivery. For example, Intelligent Science Stations is described as a mixed-reality system (Yannier et al., 2024), and IntuiGéo provides corrective feedback based on data from pen strokes on a tablet (Colliot et al., 2024). In another study, a personalised tutor robot was implemented for assessment and support, based on the Robot Operating System and using XGBoost with the robot “Duet” (Maaz et al., 2025). Programming-oriented tutoring support was present in a Scratch-integrated hint generation system that uses Whisker to check correctness and Catnip for next-step hints (Obermüller et al., 2023). These examples indicate that computational or representational methods are not merely technical implementation choices, but can shape how tutoring systems model learner knowledge, as well as how they select interventions and provide feedback. Across the systems reviewed, the choice of method used therefore appears closely tied to the system’s pedagogical function and the kind of learner support it can provide.

Lastly, some ITSs are distinguished by the inclusion of **affective, socio-emotional, reflective, or gamified support**. The adaptive cognitive load regulation framework explicitly targets metacognitive support (Zhai et al., 2025). Another adaptive learning system uses self-esteem and emotional intelligence as

learner-model inputs (Ezzaim et al., 2023); the MindfulMath Tutor provides cognitive and socio-emotional support through a pedagogical agent (Hladký et al., 2025), and PAT2Math is explicitly framed as a personal affective tutor (Morais & Jaques, 2022). Socio-cultural integration is also explicit in the interactive digital module studied by Simbolon et al. (2025). Game-based design appears in an AI-augmented learning programme that combines adaptive task sequencing, automated formative feedback, conversational agents and narrative game mechanics (Jaloliddin, 2025).

4.3. Main findings

This section synthesises the main empirical findings of the studies reviewed across the three outcome domains of the conceptual framework: learning outcomes, motivation and engagement, and socio-emotional or well-being-related effects. For ease of reading, each subsection ends with a short note on its particular policy relevance. These notes do not provide recommendations, which are formulated in Chapter 5, but highlight the educational significance of the findings in relation to interpreting the evidential base.

4.3.1. Academic performance and skill development

Of the reviewed here, 54 measure academic achievement while 33 measure skill development. Overall, the findings on both achievement and skill development are predominantly positive, although the outcomes reported are based on different types of measures.

In mathematics and related **achievement measures**, the positive effects reported mainly refer to post-test or endline mathematics performance, including standardised test scores and curriculum-aligned post-tests. These effects ranged from smaller gains, such as +0.122 standard deviations in an end-of-year mathematics test (Oreopoulos et al., 2024, Oreopoulos et al., 2026) or Hedges' $g^1 = 0.27$ for a post-treatment mathematics test given to pupils who made intensive use of an ITS (Feng et al., 2025), to larger effects such as +0.326 to +0.467 standard deviations on standardised endline outcomes and adjusted $d^2 = 0.40$ on post-test achievement, depending on the learner group, comparison

¹ Hedges' g is a [standardised effect](#) size for measuring the difference between two group means; it is a corrected version of Cohen's d , specifically adjusted to remove an upward bias that occurs when sample sizes are small (0.2 = small difference; 0.5 = medium difference; 0.8 or higher = large difference)

² d (Cohen's d) is a [standardised effect](#) size for measuring the difference between two group means (0.2 = small difference; 0.5 = medium difference; 0.8 or higher = large difference). Adjusted d refers to an adjusted dissimilarity index used in analysis to correct, for example, for small sample sizes or random chance

condition, and intensity of use (Pervaiz et al., 2025). Positive achievement findings are also reported in science, foreign languages and reading, as measured by science post-test performance, foreign-language achievement, and pre-post gains in reading-related outcomes (Yan, 2024; Yannier et al., 2024; Wijekumar et al., 2024). In one study, performance in spreadsheet skills was also higher in the experimental group than in the control group (Ezzaim et al., 2024). Where standardised metrics were available, these studies captured differences in post-test achievement or learning gains. These ranged from small effects such as eta squared³ (η^2) = 0.02 to very large effects such as $d = 1.93$ and $\eta^2 = 0.742$ (Alakayleh, 2025; Borchers et al., 2023; Shih et al., 2023).

Findings on **skill development** were likewise mainly positive and covered conceptual understanding, problem solving, reasoning, retrieval, computational thinking, digital literacy, scientific competency, self-regulated learning and reading strategy performance. Strong positive effects are reported for mathematical reasoning ($d = 1.93$) and its subdimensions (conjecture, $d = 1.81$; justification, $d = 1.83$; representation, $d = 1.85$; and metacognition, $d = 1.84$; Alakayleh, 2025), and large gains are also reported for several reading-strategy outcomes, including main-idea quality (SES = 1.07 to 1.14) and comparison-recall measures (SES = 1.07 to 1.14; Wijekumar et al., 2024). Medium-to-large gains are also reported for computational thinking (SES = 0.80; Katchapakirin et al., 2022) and goal setting ($d = 0.79$; Wang et al., 2025). Other studies describe an increase in learning progression (significantly greater learning progression by pupils in the intervention group in comparison to the control group (84.8 % vs 64.5 %; Prasertsang et al., 2025); improved autonomy - – for example, more pupils in the experimental group developed a methodology for conducting a physics experiment without substantial teacher assistance compared with those in the control group (82 % vs. 57 %; Abdulayeva et al., 2025); critical thinking – for example, pupils in the experimental group were 34 % more likely than pupils in the control group to employ methods to verify the results obtained and demonstrated a greater ability to identify logical inconsistencies in physics problem statements (Abdulayeva et al., 2025); and accuracy (Horvers et al., 2024). Process-oriented indicators also moved in a positive direction in several studies. These include more skills being practised and mastered, more lessons being passed in mathematics, and more work being completed per hour ($\beta = +0.2437$; Thomas et al., 2024b); higher practice accuracy (practice accuracy $\eta p^{24} = 0.290$; Horvers et al., 2024); and better retention after retrieval or tutoring interventions. A smaller set of studies also report positive changes in

³ eta squared (η^2) is a statistical measure of effect size used with ANOVA; values range from 0 to 1, where higher numbers mean a stronger effect.

⁴ partial eta-squared (ηp^2) is a statistical measure of effect size that indicates the proportion of variance in a dependent variable that is explained by a specific independent variable, while ignoring other variables in the model; it ranges from 0 to 1, where higher numbers mean a stronger relationship.

interest/enjoyment, perceived competence and engagement alongside performance-related outcomes (Jaloliddin, 2025; Anafa & Friday, 2025).

Several findings also indicate that **positive learning effects** were linked to the adaptive nature of the tutoring systems involved, not just to their digital format. Adaptive Bayesian knowledge tracing improved science post-test performance and increased mastery efficiency relative to random task selection, while data-driven tutor redesigns and adaptive hint policies improved sequencing, reduced over- or under-practice, or increased productive hint use (Yannier et al., 2024; Huang et al., 2021; Huang, 2025). For other systems, gains are reported in conceptual understanding, task efficiency or problem-solving completion when feedback, task sequencing, or cognitive-load regulation were adjusted to learners' states (Zhai et al., 2025; Marwiang et al., 2025; Chinjunthuk et al., 2025). These findings suggest that the effectiveness of ITSs depends not only on whether a system is labelled as being adaptive, but also on the pedagogical appropriateness and reliability of its adaptive mechanisms.

At the same time, findings also show that **effects were often conditional on implementation features, dosage, or learner characteristics**. Pervaiz et al. (2025) report stronger effects for high- than for low-dosage reading use, and stronger per-protocol than business-as-usual estimates. Wang et al. (2024) found larger effects among pupils who completed more practice. Implementation support was particularly important in those studies in which ITS use was embedded in supervised or hybrid instructional models. Hybrid models that combined human tutoring with ITS support produced stronger diagnostic gains than the use of software alone. They also increased engagement with the tutoring system and showed positive dose-response patterns, including larger gains with additional tutoring hours or improved tutor-to-student ratios (Batt et al., 2024; Thomas et al., 2024a; Thomas et al., 2024b). Similarly, the assigning of dedicated in-school supervisors substantially increased weekly platform use and produced considerable progress in mathematics, suggesting that structured supervision can transform low-intensity ITS use into sustained and productive practice (Oreopoulos et al., 2026).

Several studies also report heterogeneity by learner starting point or context, including larger gains for lower-pretest learners, those in lower grades, and for certain socioeconomic status (SES) groups. These occurred alongside differences by school type, the level of disadvantage in the school, prior knowledge, topic difficulty, and estimated grade level (Meeter, 2021; Thomas et al., 2024b). Equity-related patterns were therefore mixed: lower-achieving learners, pupils with special educational needs, low-SES learners, and girls in some flexibility-training settings appeared to benefit strongly when implementation was well supported by human tutors. In such cases these tutors could, for example, supervise classroom use, coordinate content with teachers or ensure reliable internet connectivity. However, usage was also lower or less reliable among some disadvantaged groups, indicating that access, attendance, prior achievement, infrastructure, and learners' digital skills shaped who received

sufficient dosage to benefit (Batt et al., 2024; Meeter, 2021; Olmo-Muñoz et al., 2022; Oreopoulos et al., 2026; Thomas et al., 2024a).

Some interventions also altered learning efficiency rather than just the endpoint scores – for example, by reducing the time taken to complete tasks or equations after working through examples. One study (Azevedo et al., 2020) involved the use of a step-based algebraic ITS integrated into regular classroom activities. During problem solving, the system provided immediate support through hints and feedback on errors. It also presented a worked example at the beginning of a sequence of similar equations. This improved learning efficiency, particularly in the first equation after the worked example. Another study (Zhai et al., 2025) showed that the use of an ITS that focused on optimising cognitive load, implemented in the context of classroom instruction (Grades 4 to 9), increased task completion rates by 32 %. The ITS was operationalised through a three-tier framework for cognitive load optimisation. This involved the use of real-time monitoring mechanisms, adaptive regulation algorithms and technologies such as natural language processing and multimodal data fusion to provide personalised interventions.

Despite the predominance of positive results, the effects were **not uniformly positive** across all comparisons and outcomes. Several studies report improvement over time in one or more groups but no statistically significant between-group advantage, or found no added benefit for one intervention variant over another (Bringula et al., 2025; Morais & Jaques, 2022). In addition, several studies showed that short-term improvements in the way learners worked when using the system did not always translate into broader learning gains. Some interventions increased efficiency, fluency or engagement without producing clear advantages in post-test or transfer outcomes. Others showed benefits mainly under specific conditions, such as sufficiently intensive use or in outcomes closely aligned with the intervention (Azevedo et al., 2020; Feng et al., 2025; Morais & Jaques, 2022; Wijekumar et al., 2024; Yannier et al., 2024). Explicitly weaker or negative findings were also present, including a negative low-dosage effect on Word Attack Skills (Wang et al., 2024) and slower weekly learning gains when using adaptive practice software in the post-lockdown period, compared with during school closures (Meeter, 2021). These findings highlight the importance of considering the context in which ITSs are used. For instance, the accelerated learning gains observed during school closures might stem from a strong focus on core skills – areas teachers that felt could be taught more effectively at a distance compared with other skills.

In summary, the evidence indicates that effects on learning outcomes and skill development were largely positive, but their consistency depended on intervention intensity, comparison condition, learner subgroup and the specific outcome being measured. When standardised effects are reported, these ranged from small to very large, although the robustness of these estimates was limited by variations in study design, dosage, outcome measures and reporting practices. The main educational implication is not that ITSs should automatically

be expected to improve learning, but that they appear most promising when they provide sustained, curriculum-relevant and sufficiently intensive support. In relation to discussions on policy and implementation, this means that effectiveness should be interpreted in relation to the dosage, learner group and subject matter, as well as the instructional conditions under which ITS are used. It should also be interpreted in relation to the quality of adaptive feedback, task sequencing, technical reliability and human support surrounding the system, because these factors shape whether the use of an ITS translates into measurable learning gains (Batt et al., 2024; Oreopoulos et al., 2026; Zhai et al., 2025).

4.3.2. Motivation and engagement

This section of the report addresses motivation in a broad sense. It encompasses affective-motivational outcomes such as enjoyment, hope, situational interest, self-efficacy, confidence and motivation, as well as behavioural engagement indicators such as participation, time on task, practice intensity, progression and persistence, and their links to continued system use. In total, 28 studies assessed motivation-related outcomes, while 26 studies assessed behavioural engagement.

The evidence on **motivation-related outcomes** is generally positive, although it is not uniform within all constructs and comparisons. The clearest positive results relate to enjoyment, hope, situational interest, self-efficacy, motivation, attention and confidence, with standardised effects frequently in the small-to-moderate range in the most clearly reported studies (Bettinger et al., 2020; Pervaiz et al., 2025; Uriarte-Portillo et al., 2023). For example, Uriarte-Portillo et al. (2023) show that attention and confidence, modelled as factors of motivation, increased significantly ($F^5 = 8.150$, $p^6 = 0.005$ and $F = 4.089$, $p = 0.046$) for the experimental group. Positive affective experience has also been reported under specific design conditions, such as the provision of learner choice and the use of tablet-based formats in comparison to paper-and-pencil work, although these gains did not always extend to all motivational dimensions or to differences between two active digital conditions (Clément et al., 2025; Colliot et al., 2024).

Other studies likewise report higher goal-setting, interest/enjoyment, self-efficacy, perceived competence, effort/importance, task value and learner

⁵ F-value is a number received from the statistical F-test; it indicates if the variances between groups are significantly bigger than the variance within those groups (an F-value close to 1 indicates that group means or models are likely to be equal, while a much larger F-value indicates significant differences).

⁶ p-value is a probability value that measures how likely it is that test results occur by random chance; it ranges from 0 to 1 (or 0 % to 100 %) – the lower the value, the more robust the result

interest in studying mathematics and language, with effects ranging from relatively small to clearly positive (Bettinger et al., 2020; Pervaiz et al., 2025; Wang et al., 2025). The available evidence therefore indicates that motivation is most often reported as being higher under AI-supported or digitally enhanced conditions when these involve interactive tutoring, timely feedback, learner choice or tablet-based delivery relative to comparison conditions (Clément et al., 2025; Colliot et al., 2024). However, some motivational subscales remain unchanged in otherwise positive studies, including relevance and satisfaction in one study and autonomy in another (Jaloliddin, 2025; Uriarte-Portillo et al., 2023).

The evidence also suggests that motivational effects depend on the quality of scaffolding and learner control. Features such as adaptive feedback, choice, chatbot interaction or pedagogical agents were able to support motivation and confidence, but only when they were well aligned with learners' progress and needs; poorly timed support or misaligned choice structures could instead create confusion or weaken learning outcomes (Becker et al., 2025; Clément et al., 2025; Hladký et al., 2025; Munshi et al., 2022; Su et al., 2025). However, the interpretation of motivation-related findings is limited by the fact that potential **novelty effects** are not consistently examined across the studies reviewed. Several studies involve short interventions, pilot settings, first-time use or technology-rich learning environments, in which novelty may have contributed to the positive responses on motivation-related outcomes such as enjoyment, interest, confidence, attention or affective engagement (Colliot et al., 2024; Maaz et al., 2025; Uriarte-Portillo et al., 2023). Similar concerns apply to studies using newly introduced tools or learning environments, where it remains unclear whether motivational responses observed reflect the pedagogical contribution of the system, pupils' initial interest in the technology, or both (Annuš & Kmet', 2024; Mondia et al., 2023). By contrast, studies involving repeated use or longer intervention periods offer a more suitable basis for examining whether motivational effects persist beyond initial exposure, although such studies do not necessarily isolate novelty effects directly (Morais & Jaques, 2022; Olney et al., 2025; Oreopoulos et al., 2026; Wijekumar et al., 2024).

Behavioural engagement is also usually reported as improving, but this is operationalised in multiple ways in the dataset. The strongest evidence comes from those studies that measured platform use, practice time, completed work, participation and progression. In these studies, hybrid tutoring, adaptive platforms and interventions providing feedback or recommendations increased the amount of time pupils spent using the software, as well as the number of lessons passed, weekly and total usage, the amount of practice, accuracy across settings, and the number of skills mastered per hour of platform use (Oreopoulos et al., 2026; Thomas et al., 2024a). Reported effects ranged from $d = -0.55$ to 0.35 for exercise completion in Meeter (2021) to larger differences in participation and tutor-interaction measures in Ezzaim et al. (2024) and Yu et al. (2023), respectively. The findings also show that engagement gains can include both the amount and the quality of practice, as co-regulation increased the number of problems solved as well as practice accuracy, and recommendations with

explanations increased transitions from non-engagement to active engagement (Horvers et al., 2024; Majumdar et al., 2023).

The evidence also includes modelled or correlation-based pathways which specify how motivation and engagement are linked to system use. In one study, hedonic motivation, performance expectancy, interaction with AI and willingness to communicate all had positive paths with regard to attitude towards the AI tutoring model, and attitude in turn had a strong positive path to continuous use intention (performance expectancy, $PC^7 = 0.168$; hedonic motivation, $PC = 0.318$; interaction with AI, $PC = 0.365$; willingness to communicate, $PC = 0.239$; Yan, 2024). In another study, engagement was **positively associated with both the frequency of interaction and the quality of feedback** (Alakayleh, 2025). These findings do not take the form of simple treatment-versus-control contrasts, but they still describe the reported structure of motivation and engagement in the included evidence base (Alakayleh, 2025; Yan, 2024).

At the same time, the evidence is not consistently positive and includes null, negative and heterogeneous results. The clearest heterogeneity appears in Khazanchi et al. (2024), where affective engagement favoured traditional teacher-led instruction while cognitive engagement did not differ between conditions (e.g. for mathematics achievement, Hedges' $g = 0.05$). Negative engagement-related patterns appear in the form of higher frequencies of gaming-the-system and omission errors in specific instructional units, with both patterns being associated with lower learning gains (Borchers et al., 2023). Taken together, these findings indicate that positive effects are common, but they are moderated by subgroup, condition design or the specific engagement indicator being measured (Borchers et al., 2023; Thomas et al., 2024a).

Lastly, the evidence suggests that AI-supported and digitally enhanced learning conditions tend to improve motivation and behavioural engagement, but these effects vary according to motivational constructs, engagement indicators, learner groups and implementation conditions. They should also be interpreted cautiously where studies are short, technology-rich, or do not examine whether initial motivational gains persist beyond novelty effects. From an educational perspective, the findings suggest that ITSs may support engagement when they make learning more interactive, responsive and practice-oriented. However, engagement gains do not equate to learning gains and on their own should not be treated as sufficient evidence of educational value. The evidence therefore supports a cautious interpretation: motivational and engagement benefits are most educationally meaningful when they are accompanied by sustained use, appropriate scaffolding and demonstrable progress in learning or skill development (Clément et al., 2025; Feng et al., 2025; Munshi et al., 2022; Standen et al., 2020).

⁷ The path coefficient (PC) is a standardised regression weight that measures the strength and sign of a direct causal link between two variables in a [path analysis](#) or structural equation model.

4.3.3. Well-being and other socio-emotional effects

This section examines the socio-emotional outcomes of using ITSs. It is based on the extractable findings from 13 publications included in the study set. In the studies included, socio-emotional outcomes are represented only in a minority of publications, but those studies that do report such outcomes show a favourable rather than a mixed directional pattern. The findings consistently describe either increases in positive socio-emotional indicators or decreases in negative ones, rather than a deterioration in socio-emotional functioning (Bettinger et al., 2020; Hladký et al., 2025). In those studies where emotional outcomes were measured, these are reported as being positive rather than as mixed affective responses. Emotional well-being, for example, was significantly higher during learning when learners were given a choice (Clément et al., 2025).

A related but distinct cluster concerns **attitudinal and interaction-oriented outcomes**. One study reports that attitude towards the AI tutoring model was associated positively with hedonic motivation, willingness to communicate with AI, interaction with AI, and performance expectancy, and that continuous use intention was positively associated with attitude, with attitude showing a particularly strong association with continuous use intention (Yan, 2024). Another study found a significantly higher time spent interacting with an affect-supporting pedagogical agent in the treatment condition compared with the control condition, although the magnitude of this difference is not reported in standardised form (correlation between time Interacting with the AI tutor (agent) and level of mastery (treatment group): $r = 0.28$ ($p < 0.05$); Yu et al., 2023). These findings extend the positive pattern beyond affect and motivation to attitudes towards the system and interaction with it (Yan, 2024; Yu et al., 2023). They also suggest a potential link to educational outcomes, although the strength of this link differs between studies. In Yan (2024), positive attitudes were connected mainly to continued use intention, while the same study also reports gains in language performance. In Yu et al. (2023), the connection was more direct, as increased interaction with the pedagogical agent is reported alongside higher mastery and significantly better post-test scores.

A further relevant pattern is that **reductions in cognitive load** and strain were sometimes linked to the form of support provided by the system. In a short intervention, chatbot-based support reduced intrinsic and extraneous cognitive load more than textbook-style explanations, and produced stronger affective benefits than tiered scaffolding. Meanwhile, in other studies where support consisted of adaptive quizzing and cognitive-load regulation systems, lower mental demand, effort, frustration or cognitive load are reported alongside higher engagement or task efficiency (Becker et al., 2025; Ningsih et al., 2026; Zhai et al., 2025). At the same time, affect-aware adaptation did not automatically translate into higher achievement: one system increased engagement and reduced boredom, especially for learners with intellectual disabilities, but did not produce significant short-term achievement gains (Standen et al., 2020).

A final cluster concerns reductions in **negatively valenced states and strain-related outcomes**. One study reports a lower overall state of mathematics-related anxiety, with a significant main effect of time ($\eta^2 = .357$) and a significant time-by-condition interaction ($\eta^2 = .191$) in the intervention condition compared with the control condition, indicating meaningful reductions in anxiety-related outcomes (Hladký et al., 2025). Another study reports improved cognitive efficiency as a result of reducing the overall cognitive load, as indicated by lower ratings in mental demand (3.8 vs. 5.6), required effort (4.2 vs. 6.1) and frustration (3.1 vs. 5.8) among pupils in the experimental group vs. pupils in the control group, as well as higher engagement in the experimental group than in the control group (72 % of pupils in the experimental group were classified as highly engaged, compared with only 46 % in the control group; Ningsih et al., 2026). In this part of the evidence base, the reported direction is again consistent: negative socio-emotional or strain-related experiences were reduced rather than increased (Hladký et al., 2025; Ningsih et al., 2026).

Taken together, the included studies indicate that socio-emotional outcomes are not frequently represented across the full set of publications, but when they are reported, the evidence points to a consistent positive pattern rather than to mixed socio-emotional effects. The benefits reported include positive affect, emotional well-being during learning, motivation, interest, confidence, self-efficacy, perceived competence, goal-setting, favourable attitudes towards the tutoring system, stronger interaction with an affect-supporting agent, and reductions in anxiety, frustration, effort, mental demand and cognitive load (Hladký et al., 2025; Ningsih et al., 2026; Uriarte-Portillo et al., 2023; Yan, 2024). However, because this outcome domain is covered by a comparatively small subset of studies, these findings should be read as indicative rather than conclusive. The available evidence suggests that ITSs are not just relevant to achievement, but may also affect learners' confidence, anxiety, perceived competence and emotional experience during learning.

4.4. Moderators and mechanisms

This section presents the mechanisms, as well as the moderating or contextual factors, identified in the reviewed studies. While the previous sections have focused on observed effects, the analysis that follows examines how ITSs operate pedagogically and under what conditions their effects emerge. In this sense, the section reconstructs the empirical evidence on underlying mechanisms.

4.4.1. Implementation contexts

This subsection focuses on the educational contexts, such as classroom instruction, homework, extracurricular activities, within which ITSs are implemented.

Among the included studies, the ITSs were predominantly implemented in **classroom instruction**, with 60 studies observing this setting. In many cases, the ITS formed part of regular lessons, the regular school day or the formal curriculum, and pupils used the system individually or alongside teacher support within ordinary classroom activity (Thomas et al., 2024b; Zhai et al., 2025). A recurring variation of this classroom pattern involved school-based supplemental or substitutional use, such as the partial replacement of standard subject time, use during additional support classes, high-dosage tutoring during school hours, or additional instructional periods aligned with school timetables, additional class periods and scheduled school-time sessions (Bhatt et al., 2024; Wang et al., 2024). Blended classroom arrangements are also reported, including use after instruction from a teacher, within computer labs, or in multimedia classrooms as part of structured school-based learning (Azevedo et al., 2020; Horvers et al., 2024).

Implementation outside the classroom appeared in a smaller subset of the literature reviewed. These studies described the use of ITSs in **homework, informal digital learning, summer holiday assignments and online settings** that extended beyond traditional classroom boundaries (Bettinger et al., 2020; Wang et al., 2025). Extracurricular implementation was also present, particularly in science centre and summer camp settings, and at a coding camp (Yannier et al., 2024; Yu et al., 2023). Some systems were also explicitly designed to cover classroom, homework and extracurricular phases, or to combine in-class sessions with optional home practice (Shih et al., 2023; Wen & Pan, 2025). A smaller number of studies referred only to broad educational settings without clearly specifying whether the use of ITSs occurred during classroom instruction, homework or extracurricular activity (Liu et al., 2025; Matayoshi et al., 2020).

The studies varied substantially in whether context was treated as a factor influencing the effects of the ITS. Some papers did not explicitly discuss or empirically test context, even though the intervention was situated within formal school instruction (Marwiang et al., 2025; Yan, 2024). Other papers mentioned context mainly as a boundary condition for interpretation – for example, through comments on generalisability, sample specificity or access constraints, rather than through direct analysis of contextual influence. This indicates that the presence of contextual description in the reviewed literature does not necessarily coincide with the empirical testing of context as an influencing factor (Hladký et al., 2025; Prasertsang et al., 2025b).

In those studies in which context was discussed or tested, it was often defined in terms of **institutional, demographic, cultural, geographical, linguistic and**

infrastructural conditions, rather than only through the distinction between classroom instruction, homework and extracurricular use. Several studies referred to the type of school, poverty, socioeconomic position, rural or urban location and pupil diversity as relevant contextual conditions – and some explicitly reported differences associated with these conditions, such as differences by school type or improved outcomes for lower socioeconomic groups (Uriarte-Portillo et al., 2023; Zhai et al., 2025). Context was also framed through local cultural integration, policy environments, curriculum localisation, multilingual design and national or regional conditions for implementation (Simbolon et al., 2025; Zhai et al., 2025). Simultaneously, infrastructure, digital access, teacher readiness and resource availability were repeatedly described as conditions shaping implementation or the possible scope of an ITS's effectiveness (Alenezi, 2024). Temporal disruption was another contextual strand, especially in studies discussing school closures, remote learning, pandemic-related design constraints and post-pandemic engagement (Feng et al., 2025; Meeter, 2021).

Overall, the studies show that ITSs were mainly implemented in classroom-based settings, while out-of-class use was less common and context, when considered, was typically framed in broader institutional, social, and infrastructural terms rather than only in terms of the instructional setting. As positive results are found in various contexts, both within the boundaries of the classroom and beyond, it can be assumed that context alone has no direct effect on the effectiveness of an ITS.

4.4.2. Feedback

Feedback, hints, explanations and scaffolds form one of the most direct and consistently tested design clusters. In several studies, otherwise identical platforms produced larger gains when automated feedback, hints or remedial guidance were enabled, indicating a close empirical link between visible support and learning progression (Marwiang et al., 2025; Prasertsang et al., 2025a). Marwiang et al. (2025), for example, found that improvements in the mathematical ability of pupils using an ITS that provided automated feedback exceeded the improvements of pupils using an ITS without automated feedback by 23.29 percentage points – a statistically significant margin ($t = 3.21$, $p = 0.001$, $\eta^2 = .73$). More fine-grained studies show that the precise form of feedback matters (Bai et al., 2022; Colliot et al., 2024). For example, when a word was read incorrectly on the first attempt, explicit correction cues led to greater immediate gains in reading accuracy than implicit feedback ($\beta = -1.40$, $p < .001$; Bai et al., 2022). Explanatory scaffolds, worked examples and feedback-rich learning modes were also associated with better task completion, reduced time-on-task or improved retention – although some studies suggest that such supports did not always translate into stronger comprehension or achievement. Meanwhile, visual next-step hints supported early progress and self-service without improving comprehension, and an upfront worked example improved speed rather than test performance (Matayoshi et al., 2020; Zhai et al., 2025).

This distinction is important, because ITS support for task performance does not necessarily equate to genuine or durable knowledge acquisition: some feedback designs helped learners to perform tasks more accurately, quickly or independently, but did not produce broader gains in comprehension, post-test achievement or transfer. Feedback type also mattered for **cognitive outcomes**. Diagnostic, corrective, explanatory and real-time feedback was more consistently linked to conceptual understanding, reasoning, learning progression and near-transfer outcomes than simple evaluative feedback, but these effects were often embedded within broader ITS designs and therefore could not always be isolated as independent causal mechanisms. Related evidence on cognitive load and cognitive offloading was more limited. Several studies suggest that feedback, worked examples, chatbot support or adaptive scaffolding can reduce extraneous cognitive load or support working memory, but cognitive offloading itself was rarely named or tested directly. Where offloading-like mechanisms emerged, these were usually discussed indirectly as reduced cognitive strain, automated guidance or reliance on system support; such support may be beneficial when it frees up capacity for learning, but is problematic when it encourages shallow use of the support, confusion, or gaming the system.

Borchers et al. (2023) further showed that the relationship between **feedback** and learning depended on how pupils interacted with the task. In one unit, menu-based response options combined with immediate feedback may have encouraged pupils to try out alternatives until one was accepted, rather than reasoning their way through the task. Gaming the system occurred most frequently in this unit, and the ITS tutor's within-practice learning rate was lowest. In the pen-and-paper condition, pupils could instead skip solution steps, with a greater number of these errors of omission being associated with lower learning gains. Together, these findings illustrate why feedback cannot be interpreted independently of the task and interaction design. In summary, feedback appears to be one of the mechanisms more consistently supported by ITSs, especially when it is immediate, diagnostic, explanatory and closely tied to learner errors and task demands. However, feedback should not be treated as being uniformly effective: some feedback designs mainly improved accuracy, efficiency or near-transfer outcomes, while others produced null effects, increased confusion, or were too closely embedded into broader ITS designs to be isolated as independent causal mechanisms.

4.4.3. Adaptive sequencing

The discussion below examines how adaptive sequencing is implemented in ITSs and whether adaptivity is studied primarily as a mechanism of action within the system or, less often, as a moderator of ITS effects. In total, 62 studies discuss adaptive sequencing.

In the studies included in the full-text review, the most common implementation of **adaptive sequencing** in ITSs was for real-time selection of the next task,

exercise or difficulty level on the basis of a pupil's current mastery, performance or readiness (Feng et al., 2025; Yannier et al., 2024). In such descriptions, the adaptive sequence often **guides learners towards** content described as closest to mastery, falling within the zone of proximal development, or aligned with a diagnosed proficiency level, with progression being governed by a transition from easy to difficult, or by mastery thresholds, readiness states or ongoing performance data (Feng et al., 2025; Wang et al., 2024). This indicates that immediate sequencing was the dominant form within the material. Its educational relevance lies in the possibility of aligning task difficulty and progression with pupils' current understanding, rather than requiring all learners to follow the same fixed sequence. The policy-relevant issue is therefore whether such sequencing supports appropriate challenge, sustained practice and curriculum-relevant progression. Yannier et al. (2024) examined adaptivity as a single influencing variable of ITSs, comparing an adaptive condition that included Bayesian knowledge tracing-based task selection, versus a randomised condition with tasks being chosen at random. The results showed that learning improved with adaptivity, as the average score in the post-tests (69 %) was significantly higher for the adaptive condition than for the randomised condition (54 %; $F=4.02$, $p=0.05$, $d=0.66$).

A second pattern relates the use of explicit learner models and diagnostic systems to **generate individualised pathways**. The findings describe decision trees, reinforcement learning, Bayesian knowledge tracing (Yannier et al., 2024), fuzzy inference, performance factors analysis, Elo-based estimation (Horvers et al., 2024), deep learning, XGBoost (Maaz et al., 2025), and multi-armed bandit approaches as bases for assigning pathways, recommending problems, updating next actions or recomputing activity sets (Prasertsang et al., 2025a; Uriarte-Portillo et al., 2023). The variables informing these models were not restricted to correctness; the reviewed content also referred to response times, learner preferences, learning styles, linguistic needs, emotional engagement, affective states, self-esteem, emotional intelligence, demographic factors and error patterns as inputs into adaptive sequencing (Ezzaim et al., 2024; Standen et al., 2020). In the dataset, adaptive sequencing therefore operated not only as performance-based difficulty adjustment, but also as profile-based personalisation. At the same time, adaptive sequencing was not uniformly beneficial. Some findings suggest that learners may need guidance to interpret adaptive pathways, especially when task order, difficulty or support change in ways that are not transparent to them (Ningsih et al., 2026; Verma & Kulkarni, 2025). This concern is consistent with evidence that the effects of feedback and scaffolding vary according to the design of systems, learner characteristics and task context, and that poorly calibrated support can increase confusion, overload or the shallow use of help (Borchers et al., 2023; Munshi et al., 2022; Ningsih et al., 2026). Adaptive sequencing therefore requires pedagogical framing rather than purely technical optimisation.

The included studies also distinguish between forms of adaptive sequencing according to the timing and explicitness of **instructional support**. At the

immediate level, many systems provided real-time visual feedback, automated feedback, tailored hints, remedial suggestions and scaffolding that changed as learners progressed or as particular errors appeared (Colliot et al., 2024; Ningsih et al., 2026). The support itself ranged from relatively simple step-level feedback or next-step hints to more elaborate, increasingly explicit or dialogic forms of assistance. These included tiered aids, additional rounds of scaffolding, Socratic chatbot exchanges, conversational agent, and the fading of scaffolding as competence improved (Liu et al., 2025; Olney et al., 2025). This indicates that the reviewed literature includes both simpler and more elaborate adaptive support structures, rather than a single form of adaptive sequencing.

Although immediate adaptation was most common, the material also reports **delayed or distributed forms of sequencing**. These took the form of spaced repetition, the adaptive scheduling of practice frequency, interleaving, repetition after errors, periodic reassessment and persistent-profile recommendations across sessions (Jaloliddin, 2025; Verma & Kulkarni, 2025). In addition, some implementations embedded sequencing within wider instructional structures such as tutor dashboards, optional module order and group-based task assignment systems (Oreopoulos et al., 2024; Thomas et al., 2024b). The evidence therefore points to adaptive sequencing operating according to more than one timescale: within-task, next-task, within repeated practice, and as part of broader curricular organisation.

The evidence overwhelmingly frames **adaptivity as a mechanism of action within the ITS rather than as an external moderator**. In the dataset, adaptivity was typically linked to personalised instruction, adaptive pathways, feedback, scaffolding, engagement, motivation, inclusion or learning outcomes, which places it inside the operative logic of the tutoring system itself (Feng et al., 2025; Wang et al., 2024). Explicit moderator formulations were rare and were commonly limited to statements about attention from tutors, learner type and socioeconomic background, interaction patterns and teacher involvement, or to prior achievement and intensity of usage (Thomas et al., 2024b; Zhai et al., 2025). Two additional papers described adaptivity playing a more tentative moderator-like role, through explanatory prompts that shaped engagement or through comparison with non-adaptive practice contexts (Majumdar et al., 2023; Meeter, 2021). Majumdar et al. (2023), for example, found that a small set of papers described adaptive or partly adaptive features, but did not explicitly investigate adaptivity as either a mechanism or a moderator, or noted that the sequence of focus was fixed or only partially specified (Olmo-Muñoz et al., 2022; Oreopoulos et al., 2026).

Taken together, the dataset describes adaptive sequencing in ITSs primarily as the immediate, performance- or model-based selection of content, difficulty and support, with additional evidence for delayed and distributed sequencing through spacing, reassessment and persistent learner profiles across sessions. The sequencing of support ranges from simple hints or step-level feedback to more elaborate scaffolded and dialogic assistance, while the evidentiary framing is

predominantly mechanistic rather than moderational (Feng et al., 2025; Olney et al., 2025; Thomas et al., 2024b; Verma & Kulkarni, 2025). The main conclusion is not that a specific sequencing method is superior, but that sequencing is educationally relevant because it shapes the match between pupils' current understanding, the tasks they receive, and the support provided during learning. However, the studies reviewed rarely isolate adaptive sequencing as an independent causal factor. Its effects should therefore be interpreted as part of broader ITS designs that combine sequencing with feedback, scaffolding, curriculum alignment, teacher involvement and sustained use.

4.4.4. Scaffolding

All 65 of the included studies examined scaffolding. Scaffolding was most commonly described as **direct cognitive support** in the form of hints, explanations, worked examples, prompts, feedback and step-by-step guidance (Wutthisarn & Junpeng, 2024; Zhai et al., 2025). **Guided practice** was also a recurring element, and appeared through warm-up tasks, practice-question sessions, readiness-based topic progression, repeated practice, guided tutoring sessions and structured problem-solving sequences (Bhatt et al., 2024; Matayoshi et al., 2020). In several papers, these forms of support were supplemented by explicit error correction and explanatory feedback, including real-time visual cues, explanatory feedback, guidance and concept visualisations (Colliot et al., 2024; Yannier et al., 2024). This indicates that scaffolding was not a marginal feature of the ITSs reviewed, but a recurring design logic through which systems attempted to make tasks manageable, provide contingent support, and guide learners through problem solving.

A second recurrent theme concerns the **adaptive organisation of support**. Many systems adjusted the selection of questions, task difficulty, content sequencing or recommended exercises in response to prior performance, attempts, knowledge gaps, affective state or recent answers (Pervaiz et al., 2025; Simbolon et al., 2025). These adaptive processes were sometimes embedded into dashboards, recommendation systems or teacher-facing monitoring structures that helped them to direct attention towards pupils who needed differentiated support (Oreopoulos et al., 2024; Thomas et al., 2024b). The extracted findings also show that scaffolding was not limited to cognitive guidance. Several papers describe metacognitive, motivational and socio-emotional support, including support for self-regulation, goal-setting or metacognitive prompts, encouragement, and relationship-building human tutoring (Su et al., 2025; Thomas et al., 2024b). Scaffolding therefore appeared both as learner-facing support and as part of the broader instructional infrastructure around ITSs, including teacher dashboards, adaptive recommendations and human guidance.

A smaller subset of studies explicitly analysed scaffolding as an effective factor. Their findings indicate that **contextual support**, delivered through natural

language processing, addressed comprehension gaps and was described as being important for cognitive-load management and learning outcomes, while **dynamic composition scaffolding** and explicit practice on hard knowledge components were associated with improved learning outcomes and more effective practice experiences (Huang et al., 2021; Zhai et al., 2025). For example, in Zhai et al. (2025), an intelligent cognitive-load regulation system improved the academic scores of pupils from lower-SES backgrounds, increased task-completion rates by 32 %, and reduced reported cognitive load by 23.1 % on average.

Worked examples have also been empirically examined as scaffolding, and are reported to reduce cognitive load and improve learning efficiency – especially immediately after the example had been worked through (Azevedo et al., 2020). Other empirical studies linked scaffolding to speaking performance, vocabulary, goal setting, self-evaluation, motivation, learning outcomes and self-regulated learning behaviours (Munshi et al., 2022; Wang et al., 2025). However, the evidence base is uneven: while scaffolding-related features appeared across many systems, only a smaller set of studies treat scaffolding itself as the focal mechanism or compared specific scaffolding variants. In many other studies, scaffolding is bundled with feedback, adaptive sequencing, recommendation systems, dashboards, human tutoring, or broader personalised instruction, making it difficult to determine whether observed effects were caused by scaffolding specifically or by the wider ITS package.

The dataset also contains more qualified findings. Borchers et al. (2023) found that when compared with paper-based practice, ITSs provided more than twice as many practice opportunities that eventually resulted in correct answers – but this did not result in greater overall learning gains. Different types of scaffolding were used for higher- and lower-performing learners (Liu et al., 2025), and excessive scaffolding was described as a possible source of cognitive overload, with gradual onboarding suggested as a response (Ningsih et al., 2026). These findings suggest that scaffolding is not automatically beneficial. Its effects appear to depend on timing, calibration, learner proficiency, task difficulty and whether support encourages active reasoning rather than the shallow use of help or reliance on guidance from the system.

This cautious interpretation is supported by evidence on cognitive load, but this does not establish the existence of long-term effects on pupils' cognitive capacity. Some studies report that scaffolding, worked examples, chatbot support or adaptive quizzing reduced cognitive load or improved learning efficiency, but the evidence reviewed does not identify studies that have directly examined whether such reductions produced longer-term improvements in cognitive capacity. Azevedo et al. (2020), for example, found efficiency gains after worked examples, but these related to a reduction in solving time rather than long-term changes in cognitive capacity. Similarly, studies of scaffolding and self-regulated learning, such as Munshi et al. (2022) and Wang et al. (2025), examined learning processes over time but did not specifically test long-term cognitive-load

reduction or cognitive-capacity outcomes. The closest longer-duration evidence comes from Zhai et al. (2025). This study reports improved academic performance and task completion, lower cognitive load and self-regulated learning over a 6-month period, although true long-term effects remained unestablished.

Taken together, the dataset shows that while hints, guided practice, adaptive sequencing and human- or agent-based support are widely reported as scaffolding elements in ITSs, explicit testing of scaffolding as a distinct factor occurred less frequently and was concentrated in studies of adaptive, dialogic, difficulty-based or worked example-based support (Azevedo et al., 2020; Wang et al., 2025). The main conclusion is therefore not simply that scaffolding is a common feature of ITSs, but that it is theoretically central and empirically under-isolated. While scaffolding appears to be a core mechanism through which ITSs provide contingent support, the available studies often do not separate its effects from those produced by feedback, adaptivity, teacher involvement or broader system design. This limits the ability of the field to specify which scaffolding components work best, for whom, and under what conditions.

4.4.5. Technological design features

Throughout the studies on ITSs reviewed, the technological design features described are clustered around dashboards and progress displays, conversational or agentic interfaces, gamification and rewards, as well as multimedia or augmented reality formats. The relationship between these features and the outcomes of the use of ITSs was tested using several empirical strategies. These include randomised and quasi-experimental comparisons of alternative interfaces, dose-response or usage analyses, correlational models, and post-use perception studies (Oreopoulos et al., 2024; Thomas et al., 2024b). At the same time, a substantial portion of the evidence comes from evaluations of complete bundles of features rather than isolated elements of interfaces, which means that many findings concern the effectiveness of integrated systems rather than individual functionalities (Verma & Kulkarni, 2025; Wijekumar et al., 2024).

Dashboards, mastery indicators, reports and progress visualisations constitute a major cluster of features. In hybrid tutoring settings, real-time tutor dashboards were used to prioritise assistance and differentiate support, and gains were more strongly associated with mastery-linked process indicators than with exposure time alone (Oreopoulos et al., 2024; Thomas et al., 2024b). In Thomas et al. (2024b), an implementation that combined a real-time tutor dashboard with a lower human-tutor-to-pupil ratio was associated with approximately 0.36 additional completed MATHia workspaces per hour. However, because both dashboard availability and the tutor-to-pupil ratio differed between the compared implementation settings, this estimate cannot be attributed to the dashboard alone.

Learner-facing dashboards, knowledge maps and recommendation panels were linked to self-monitoring, perceived personalisation or to increases in active engagement, but their relationship to learning was more mixed. Recommendation dashboards were more effective when paired with explanations, and progress displays sometimes improved regulation or user ratings without producing clear advantages in terms of overall achievement. In Majumdar et al. (2023), pupils receiving recommendations with explanations achieved a mean score of approximately 63.1, compared with 52.2 among pupils who did not use the recommendation system ($\eta^2 \approx 0.048$). A third group, which received recommendations without explanations, achieved a mean score of approximately 59.0 and did not differ significantly from either comparison group. The material included in this review therefore suggests that while dashboards are consistently used as visibility and coordination mechanisms, their contribution to achievement depends on whether or not they were linked to feedback that informs subsequent learning or instructional decisions, mastery progression or explanatory features.

Conversational agents, avatars and social-emotional interfaces form another prominent theme in the reviewed texts. Compared with less interactive or more traditional formats, conversational and adaptive agents were linked to lower cognitive load, more favourable attitudes and motivational outcomes, higher vocabulary or post-test performance, and stronger engagement (Yan, 2024; Wang et al., 2025). The interfaces examined included chatbots, animated pedagogical agents, talking-head tutors and robot tutors, often combined with immediate guidance, further questioning or personalised feedback. In Wang et al. (2025), an interactive agent providing personalised feedback and adaptive prompts produced higher vocabulary outcomes than a fixed, non-interactive version of the same agent ($d \approx 0.36$), as well as a larger difference in goal setting ($d \approx 0.79$). Differences in self-evaluation ($d \approx 0.43$) and motivation ($d \approx 0.22$) were positive, but were not statistically significant. Social-emotional extensions, however, showed a more selective pattern. A mindfulness-oriented agent reduced mathematics anxiety without improving performance, and supportive avatar or empathy features were rated positively or associated with positive outcomes mainly under higher usage or in descriptive accounts (Feng et al., 2025; Hladký et al., 2025). Added social presence did not automatically improve achievement, optional task prompts delivered by avatars did not outperform text-only prompts, and encouragements given by one agent-based science tutor did not produce measurable affective benefits (Munshi et al., 2022; Su et al., 2025).

Gamification, rewards, narrative framing and playful choice architectures were described in the reviewed literature, but the pattern of outcomes associated with these was conditional rather than uniformly positive. Badges, medals, trophies, streaks, points, leaderboards, narrative quests and other reward structures were often linked to motivation, enjoyment or perceived engagement, but direct comparative evidence indicates that these features did not automatically lead to better learning outcomes. In one comparison, adaptive multimodal presentation outperformed a gamified interface under the same adaptive navigation and feedback structure; in another, a choice-based

gamification element improved learning only when it was embedded into an adaptive personalised sequence (Clément et al., 2025; Sayed et al., 2022). In Sayed et al. (2022), scores increased by approximately 8.9 % when using adaptive multimodal presentation, and by 6.6 % using the gamified interface. Learning efficiency increased by approximately 169 % and 150 %, respectively. In Clément et al. (2025), the condition that combined personalised sequencing with learner choice produced learning gains of approximately 3 points, whereas adding the same choice feature to a fixed linear sequence produced gains of only approximately 1–1.5 points. These study-specific estimates reinforce the finding that the effects of gamification depend on the instructional sequence in which it is embedded.

Multimedia, augmented reality, mixed-reality and input-modality features show a similarly conditional pattern. Adaptive augmented reality and interactive multimedia systems were linked to higher learning and motivation when combined with adaptive guidance or feedback, whereas changes in input modality affected efficiency, usability or near-transfer performance more often than they did broader learning outcomes (Colliot et al., 2024; Morais & Jaques, 2022; Simbolon et al., 2025; Uriarte-Portillo et al., 2023).

The studies reviewed also show that workflow design and interface constraints can shape outcomes. Same-session enforcement of different solution paths improved problem-solving proficiency, and learning-mode retrieval outperformed assessment-mode retrieval for later retention (Matayoshi et al., 2020; Olmo-Muñoz et al., 2022). Other workflow-oriented features, such as anti-guessing mechanisms, host-controlled settings or recommendation mechanisms produced clearer changes in interaction patterns than in post-test outcomes, indicating that some design elements operate primarily on learner behaviour rather than directly on achievement (Bringula et al., 2025; Huang, 2025).

Taken together, the evidence supports a differentiated conclusion: in the studies, dashboards and conversational agents often supported regulation, engagement or affect, while gamification, multimedia and workflow-related features showed benefits that were more conditional or selective within broader instructional systems (Clément et al., 2025; Olmo-Muñoz et al., 2022; Thomas et al., 2024b; Uriarte-Portillo et al., 2023).

4.4.6. Educational stage and age group

The analysis in this subsection addresses whether differences between age groups in the effects of ITSs have been empirically tested and, from an educational perspective, what any differences reported suggest with regard to the school stages at which support from ITSs appears more or less beneficial, drawing on findings from five studies (see also subsection 4.1.4 for a graphic illustrating age distribution). These five studies were conducted within various national education systems. Grade labels are therefore retained as reported in

the original publications and should not be interpreted as being directly equivalent between countries. Where available, pupils' ages and the broader educational stages used by the respective authors are reported alongside these labels to support interpretation.

The evidence applies mainly to primary and lower-secondary educational levels, with age ranges reported only in some studies. One study focused on Grade 5 pupils aged 10 to 13 years (Horvers et al., 2024), while another compared third-graders aged 8 to 9 years and fourth graders aged 9 to 10 years (Jaloliddin, 2025). Other studies report findings within broader school levels, including Grades 3–6, Grades 7–8, or lower versus higher grades without specifying exact ages (Meeter, 2021; Oreopoulos et al., 2024). In summary, the evidence base is concentrated in school-age populations and is reported mainly in terms of grade level, with age information included less consistently.

A first issue deals with differences in the effectiveness of ITS use between **grade levels**. Several findings indicate stronger effects on younger or lower-grade pupils. However, this pattern is not fully uniform. Wijekumar et al. (2020) report that fifth-grade pupils showed higher effect sizes than fourth-grade pupils, indicating that variation can occur even between adjacent primary grades. This finding does not support a simple linear pattern in which younger pupils benefit more. On a broader level, Meeter (2021) states that adaptive practice software had a stronger impact on learning in lower grades (such as 4 and 5) than in higher grades. Similarly, Oreopoulos et al. (2024) report significant effects in Grades 3–6, but no significant effects in Grades 7–8, which points to a difference between primary and later grade levels. Taken together, these studies suggest stronger effects in lower grades, although the level of detail varies between publications. More importantly, they indicate that educational stage is not merely a background characteristic, but helps to explain differences in the effectiveness of ITSs. Across the evidence reviewed, stronger results in lower grades coincide with two recurring features: first, learners at different school stages appear to differ in the self-regulatory and metacognitive demands that ITSs place upon them (Horvers et al., 2024; Jaloliddin, 2025); second, school stages also differ in implementation conditions, especially the amount of structured practice time and curriculum flexibility available for sustained use (Meeter, 2021; Oreopoulos et al., 2024). In this sense, differences between primary and later grades are best understood not as isolated age-related effects, but as differences in developmental readiness and instructional opportunity.

A second issue deals with variation within primary-level settings and the limits of **subgroup differentiation**. Although Oreopoulos et al. (2024) report effects for Grades 3–6, subgroup analyses within primary school classrooms did not show strong patterns of differential benefits across age groups among this range. This contrasts with studies that report more specific differences between nearby grades, such as the higher effect sizes for Grade 5 than Grade 4 reported by Wijekumar et al. (2020), and the differentiated pattern between third- and fourth-graders reported by Jaloliddin (2025). The evidence therefore includes both

variation within primary levels as well as findings that suggest limited subgroup differentiation among that same range. This implies that primary education should not be treated as a fully homogeneous stage when interpreting the effects of ITSs.

A third issue deals with **differences in the types of outcomes** observed across age groups. Jaloliddin (2025) does not report a single uniform advantage for one level, but instead describes a differentiated pattern in which third-graders showed stronger gains in enjoyment, whereas fourth-graders showed steadier improvements in competence ratings. This indicates that reported age-group differences were not limited to overall effectiveness, but also varied according to the outcome domain considered. The clearest evidence for this comes from Jaloliddin (2025), who reports a differentiated pattern across motivational outcomes, with stronger gains in enjoyment for third-graders and steadier improvements in perceived competence for fourth-graders.

A fourth issue deals with the **evidence on level differences** – specifically, whether age or grade effects were tested directly, detailed descriptively, or were only reflected in sample characteristics. Some studies (Oreopoulos et al., 2024; Wijekumar et al., 2020) explicitly examine differential effects between levels. Oreopoulos et al. (2024) report a significant intent-to-treat effect of 0.17 for Grades 3–6, but no significant effect for Grades 7–8. Wijekumar et al. (2020) report small but meaningful effects for fourth-grade pupils and moderate-to-large effects for fifth-grade pupils. The direct comparative evidence therefore points to significant effects in Grades 3–6, with the highest effects being seen in the fifth grade within the comparison conducted by Wijekumar et al. (2020). This evidence does not, however, permit a general ranking across the studies. Meeter (2021) also reports a stronger impact in lower grades, but does not provide explicit age ranges or a detailed analysis beyond grade-level differences. Horvers et al. (2024), in contrast, primarily identifies the applicable age and grade range and notes that the control group was 6 months older than the experimental group, while also stating that the study did not provide detailed empirical evidence on age-group differences with regard to effects. Meeter (2021) and Horvers et al. (2024) therefore provide contextual information but no additional direct evidence on differential effectiveness between age- or grade-group differences, descriptive grade-related patterns, and studies that mainly report sample characteristics. This is an important consideration when interpreting the evidence, because not all reported age- or level-related patterns are based on equally strong empirical tests.

In summary, the evidence on age or educational stage applies mainly to Grades 3–8, with the clearest detail concentrated in the primary school range (in some cases, linked to specific ages). Reported differences between age groups or educational levels were present in several publications, especially in comparisons between lower and higher grades or between adjacent primary grades. However, not all studies provided detailed age-based analyses, and one study reports no strong subgroup pattern within primary classrooms despite

broader differences between school levels (Horvers et al., 2024; Oreopoulos et al., 2024). Overall, the most relevant conclusion educationally is therefore not simply that “younger pupils benefit more”, but that the effects of ITSs appear to vary between school stages, adjacent grades and outcome domains, with the clearest and most differentiated evidence currently concentrated in the primary years.

4.4.7. Subject domain

The included studies indicate that ITSs are used in a **broad range of academic disciplines**. The discipline most frequently represented is mathematics, including arithmetic, algebra, equivalent fractions, the multiplication and division of fractions, linear graphs, numbers and algebra, and measurement and geometry (Annuš & Kmet', 2024; Prasertsang et al., 2025b; Shih et al., 2021, 2023; Thampitak & Junpeng, 2025; Thomas et al., 2024b). Reports on the use of ITSs are also emerging in language-related disciplines, including foreign language education, English language learning, language education at primary level, Bahasa proficiency, reading education, and reading comprehension (Liu et al., 2025; Ningsih et al., 2026; Wang et al., 2024; Wang et al. 2025; Yan, 2024). Science-related applications include general science, middle school science, physics, biology and science literacy topics such as climate change, energy, motion and biodiversity (Abdulayeva et al., 2025; Simbolon et al., 2025). Additional uses are reported in computer science, programming education and computational thinking, as well as in broader general education and special education contexts covering multiple domains such as language, mathematics and social skills (Ezzaim et al., 2024; Katchapakirin et al., 2022; Obermüller et al., 2023; Standen et al., 2020). The results show that the use of ITSs appears to lead to an improvement in learning outcomes, with effect sizes in outcome measures differing between subjects (see subsection 3.3.1). Some studies also cover more than one subject area, such as mathematics and language, natural sciences and mathematics, or computer science and French as a foreign language (Bettinger et al., 2020; Ezzaim et al., 2023). Ezzaim et al. (2023), for example, found for the experimental group a mean difference in computer science of +3.25 points (15.78 vs 12.53) and in French as a foreign language a mean difference of +3.31 points (13.78 vs 10.47) compared with the control group, indicating only minor differences in a between-subject comparison.

Cross-subject and cross-domain differences in reported ITS effects

The material includes several cases in which differences between subjects or domains have been explicitly reported. In a general education context, performance indicators differed across mathematics problem-solving, language learning, science experiments, collaborative projects, self-paced review and exam preparation, with higher outcomes being reported for self-paced review and mathematics problem-solving than for science experiments and collaborative

projects (Zhai et al., 2025). A direct contrast between mathematics and language is also reported in computer-assisted learning, where positive effects were found for mathematics at both dosage levels, whereas language showed diminishing returns at higher dosage levels (Bettinger et al., 2020). In reading comprehension, Mondia et al. (2023) report significant improvements in learning outcomes. Reading-comprehension scores in the experimental group increased by 13.9 % from pre-test to post-test, while the control group showed a decline. This percentage refers to the study-specific score and is not a standardised effect size.⁸ Multi-subject evidence is also reported in a study covering computer science and French as a foreign language, in which both subjects showed higher scores in the experimental group than in the control group, although the score patterns differed between the two subjects (Ezzaim et al., 2023). These findings indicate that subject area was not treated uniformly in the dataset and that differences in outcomes between subjects have been explicitly noted in several studies.

Mathematics as the most extensively represented discipline

Mathematics is the subject area in which ITS applications are most heavily represented in the studies examined. In mathematics, positive effects are reported for automated feedback in numbers and algebra and in measurement and geometry, where experimental groups showed stronger procedural, conceptual, or progression outcomes than control groups (Marwiang et al., 2025; Prasertsang et al., 2025b; Thampitak & Junpeng, 2025). Marwiang et al. (2025) report an effect size of $\eta^2 = 0.73$, with the experimental group reaching an advanced learning-progression level of 84.83 %, compared with 61.54 % for the control group. Similarly, Prasertsang et al. (2025b) report average learning progression of 84.83 % in the experimental group vs. 64.54 % in the control group. In measurement and geometry, Thampitak and Junpeng (2025) report progression scores of 86.29 % and 69.89 %, respectively.

Positive results are also reported for hybrid human–artificial intelligence tutoring, one-on-one dialogue-based tutoring and intelligent game-based or adaptive systems that improved proficiency, practice behaviour, methodical problem solving or the pace of progress (Bringula et al., 2025; Meeter, 2021; Shih et al., 2021; Thomas et al., 2024b). At the same time, mathematics findings also show that the effects of ITS use were not uniform across all mathematical content. In linear graphs, ITSs produced more eventually correct practice opportunities than paper practice, but did not result in stronger overall learning gains, and effectiveness differed by curricular unit, favouring the ITS for the graphing of linear equations and paper for qualitative interpretation of graphs (Borchers et al., 2023). Similarly, learning gains were comparable for easy and intermediate topics

⁸ Raw score differences and percentage changes refer to the outcome scales used in the respective studies.

but differed for a more difficult topic, where the control condition outperformed the ITS condition (Horvers et al., 2024). Therefore, the studies present mathematics not as a single homogeneous discipline, but as a set of subdomains and task types in which the reported effects of ITSs differ.

Language, reading and literacy-related uses

Outcomes of ITS use in language and reading disciplines were also diverse. In English language learning, ITS-supported instruction was associated with greater vocabulary gains, higher goal-setting scores, and a less sharp decline in some self-regulated learning indicators over time (Wang et al., 2025). The interactive condition produced an effect of $d \approx 0.36$ for vocabulary and $d \approx 0.79$ for goal setting. Differences in self-evaluation ($d \approx 0.43$) and motivation ($d \approx 0.22$) were positive but not statistically significant.

In foreign language education, ITSs improved all dimensions of proficiency, while differences are reported by grade level but not by gender (Yan, 2024). In primary English learning, third graders showed stronger gains in enjoyment, whereas fourth graders showed steadier improvements in competence ratings (Jaloliddin, 2025). Reading-related studies further show that the effects of ITSs could vary by reading component or design feature. Higher attendance was associated with better outcomes in passage comprehension and word attack (Wang et al., 2024). Explicit versus implicit feedback in first-grade reading produced different effects depending on word context, initial reading speed and accuracy, and word length, with explicit feedback improving accuracy more for initially incorrect words, while both feedback types involved a trade-off between accuracy and speed (Bai et al., 2022). In content-area reading comprehension, the effects of ITS use differed by grade level, with small but meaningful effects in fourth grade and moderate-to-large effects in fifth grade (Wijekumar et al., 2019). These findings show that language and literacy-related ITS applications have been reported in foreign language learning, English learning and reading comprehension, and that differences in their outcomes are described in relation to grade, task, and feedback configuration.

Applications in science and physics

The studies covering science-related use of ITSs report positive effects in several disciplines and subdomains. In science literacy, an AI-driven digital module improved understanding across content areas such as matter, climate change, energy, motion and biodiversity, with the experimental group showing a substantially larger increase in scores than the control group (Simbolon et al., 2025). In the development of science competency, an ITS improved the explanation of scientific phenomena, the evaluation and design of scientific enquiry, and the interpretation of data and use of evidence – although the magnitude of these improvements differed between these three dimensions

(Wutthisarn & Junpeng, 2024). Studies concerning physics report gains in academic performance, practical skills and research competencies while emphasising the need to tailor ITSs to the instructional characteristics of the discipline (Yannier et al., 2024; Abdulayeva et al., 2025). In relation to early physics principles, adaptivity based on Bayesian knowledge tracing was described as being effective in science learning (Yannier et al., 2024). In Abdulayeva et al. (2025), 76 % of pupils in the experimental group successfully solved advanced problems, compared with 52 % in the control group. To sum up the above findings, these studies show that the use of ITSs in science-related disciplines encompasses both cognitive and affective outcomes, which may vary between subdomains and competency dimensions.

Computer science, programming and computational thinking

ITSs have also been used in computer science, programming education and computational thinking. In the high school computer science curriculum, ITSs have been associated with improved academic performance and engagement (Ezzaim et al., 2024). In this study, the experimental group achieved an average academic score of 15.94, compared with 11.92 in the control group, while participation rates were 93.58 % and 60.00 %, respectively. These values refer to the study-specific assessment and engagement measures. ScratchThAI is reported to have improved educational outcomes and engagement in computational thinking (Katchapakirin et al., 2022). In programming education, visual next-step hints improved speed and reduced help requests at lower complexity levels, but the quality of hints deteriorated as program complexity increased (Obermüller et al., 2023). These findings indicate that in computer science-related domains, ITSs were used not only to improve performance and engagement, but also in ways that were sensitive to the complexity of programming tasks.

Reported heterogeneity within disciplines

A substantial portion of the studies describe variation in the effects of ITSs *within* disciplines, rather than just *between* disciplines. In mathematics, lower-achieving pupils benefitted more from hybrid tutoring, with ITS use proving to be particularly effective in supporting learning about fractions for lower-achieving pupils, while no significant differences were observed among high-achieving pupils (Shih et al., 2023; Thomas et al., 2024b). In addition, studies covering mathematics report differences associated with gender, instructional flexibility, grade level and device or input configuration, although not all of these are framed as cross-subject contrasts (Meeter, 2021; Morais & Jaques, 2022; Olmo-Muñoz et al., 2022). In science, high-performing pupils used feedback more effectively than low-performing pupils, and strategic hints were more useful than encouragement prompts (Munshi et al., 2022). In language-related studies, differences are reported by grade level, vocabulary learning and trajectories of self-regulated

learning variables (Wang et al., 2025; Yan, 2024). In the contexts of general and special education, learning achievement and engagement are reported to differ according to intellectual disability level and autism spectrum condition subgroup (Standen et al., 2020). Therefore, studies frequently report heterogeneity in the effects of ITS use within individual subject areas, including differences by learner group, grade and topic.

Single-subject studies

Not all studies explicitly reported discipline-specific or subject-specific differences. Several studies were restricted to a single subject and reported effectiveness in that subject without making cross-subject comparisons or explicit subject-specific claims. These include studies on mathematics, algebra, language learning and computational thinking (Huang et al., 2020; Ningsih et al., 2026; Verma & Kulkarni, 2025). Other studies explicitly state that subject-specific differences were not reported or tested, even though ITSs were used in the context of a discipline, such as for mathematics achievement and engagement, or that results were consistent across learner characteristics rather than being differentiated by subject-specific conditions (Clément et al., 2025; Khazanchi et al., 2024). This indicates that the reviewed literature contains both studies that explicitly examined or described differential effects, and studies that were discipline-specific in application but not comparative in their treatment of subject differences.

Scope of ITS use across disciplines and subject-related differences

The studies reviewed examine the use of ITSs in mathematics, language and foreign language education, reading and reading comprehension, science, physics, computer science, programming education, computational thinking, geometry, algebra and broader general and special education contexts that range across multiple subject areas (Simbolon et al., 2025; Thomas et al., 2024b; Wang et al., 2024). Other studies also report the use of ITSs in computer science and broader general and special education contexts (Ezzaim et al., 2024; Standen et al., 2020). Discipline-specific or subject-related differences in effects have been reported in multiple ways: through direct contrasts between subjects, through differences between subdomains or curricular units within a subject, and through variations between competency dimensions within discipline-specific ITS implementations (Bettinger et al., 2020; Borchers et al., 2023; Wutthisarn & Junpeng, 2024). At the same time, a subset of the studies did not explicitly test or report such differences and instead remain confined to single-subject evaluations of effectiveness (Clément et al., 2025; Khazanchi et al., 2024; Verma & Kulkarni, 2025).

4.4.8. Role of teachers and parental involvement

This subsection outlines how teachers and parents have been described in the implementation of ITSs, and whether their roles were examined only descriptively or also as factors that influence outcomes.

Role of teachers in ITS implementation

Although many of the activities described in the studies resemble established teaching practices, the specific role that teachers play in relation to ITSs lies in **interpreting system-generated information** and integrating automated support into classroom instruction. Across the reviewed studies, teachers used dashboards, reports and behavioural traces to identify learning needs, allocate attention, assign additional support and adapt instruction (Simbolon et al., 2025; Wijekumar et al., 2024). They therefore acted as mediators between the system and the learner by determining how automated recommendations were used, when additional human support was required, and how use of the system aligned with curricular goals. They were also described as **providing guidance**, scaffolding, socio-motivational support, relationship building and goal-oriented feedback rather than simply supervising the use of the technology (Alakayleh, 2025; Thomas et al., 2024a). Some studies also described teachers as **contributing to system design**, content creation, assessment development, validation and curricular alignment (Yan, 2024). Some papers describe narrower functions for teachers, such as selecting participants, providing classroom access or enabling study procedures (Bringula et al., 2025; Clément et al., 2025). These findings suggest that ITSs do not remove the need for teacher judgement, but change the information that is available for pedagogical decision-making and classroom orchestration.

Teacher training and implementation conditions

Beyond their pedagogical role, teachers also assumed **operational responsibility** for sustaining the use of ITSs. This included coordinating sessions, sending reminders, troubleshooting technical problems and maintaining the fidelity of the implementation (Pervaiz et al., 2025; Wang et al., 2024). This indicates that the implementation of ITSs required ongoing organisational support rather than a one-off technical introduction. Training, professional development, collaborative support and stable routines were repeatedly reported as conditions that supported implementation, particularly where teachers had to use dashboards, monitor progress, maintain classroom routines or integrate the ITS into curriculum-aligned instruction (Pervaiz et al., 2025; Wijekumar et al., 2024). However, the studies reviewed rarely examined teacher training as an independent causal factor. Where training was described, this ranged from brief technical familiarisation to more structured professional development, coaching, office hours, fidelity checks, dashboard training and

professional learning communities, but its separate effect on student outcomes was usually not isolated (Feng et al., 2025; Pervaiz et al., 2025; Wijekumar et al., 2024). Therefore, the evidence suggests that teacher training is relevant not only for basic technical use, but also for interpreting dashboards, using learner data, integrating ITSs into curriculum-aligned instruction, and maintaining consistent classroom routines. In this sense, the primary importance of teacher training appears to be in relation to implementation quality: training supports teachers' ability to act on system recommendations, to troubleshoot technical issues, to sustain student practice, and to align the use of the ITS with classroom instruction, rather than it operating as a separately tested driver of achievement gains (Oreopoulos et al., 2024; Oreopoulos et al., 2026; Pervaiz et al., 2025).

In relation to this, teacher workload, limited time, limited digital or subject-matter competence, weak buy-in and infrastructure problems were described as **implementation constraints** (Oreopoulos et al., 2026). In addition, teacher training alone appeared insufficient where organisational capacity, scheduling, infrastructure or buy-in were weak; ongoing support and dedicated implementation roles were more closely associated with sustained use and stronger outcomes (Oreopoulos et al., 2024; Oreopoulos et al., 2026). This indicates that teacher capacity and professional development are important implementation conditions for ITSs, especially where systems require teachers to monitor progress, act on recommendations or combine automated support with direct instruction. Several studies also describe the **technology as reducing manual monitoring**, grading or administrative work – thereby freeing up teachers' time for individualised support or instructional design (Ningsih et al., 2026; Verma & Kulkarni, 2025). A smaller group of papers describe teacher involvement as being intentionally limited or largely technical, logistical or access-related (Olmo-Muñoz et al., 2022; Wang et al., 2025).

Parental involvement in the implementation of ITSs

Many papers describe the involvement of parent in procedural terms and as being limited to informed consent, passive consent, opt-out procedures or permission for participation or data use (Feng et al., 2025; Uriarte-Portillo et al., 2023). More active parental roles were reported less frequently, and took the form of feedback on affordability and children's adaptation, home guidance and additional support, extending learning outside the classroom, helping children to understand and use the system, or using assessment reports to support children's development (Alenezi, 2024; Zhai et al., 2025). These findings are directly relevant to ITSs only in cases where parents supported access, home use, the interpretation of progress information, or the acceptance of the technology. Broader parental characteristics were reported only occasionally and were rarely connected to the specific adaptive functions of the systems (Meeter, 2021; Wutthisarn & Junpeng, 2024).

Evidence on the roles of adults as implementation factors

In the included studies, the **roles of adults are more often described as implementation conditions** than empirically tested as determinants of outcomes (Pervaiz et al., 2025; Wang et al., 2025). Where teacher influence has been examined, the findings reported link socio-motivational and data-informed support to **higher engagement and achievement**. In Thomas et al. (2024a), a hybrid condition combining adaptive software with dashboard-guided human tutoring produced larger increases in diagnostic scores than software-only use among general-education pupils (4.2 % vs. 2.6 %) and pupils receiving special education (6.3% vs. 3.8%). However, these figures apply to the hybrid model as a whole, and do not isolate the effect of the human tutor or dashboard. Other findings link larger benefits for pupils who scored higher in the pre-test, teacher buy-in, fidelity, high expectations, monitoring and commitment to higher practice time and improved outcomes, whereas limited capacity or buy-in is linked to low usage and minimal learning gains (Oreopoulos et al., 2024; Thomas et al., 2024b). These findings reinforce the interpretation that teacher involvement matters most when it shapes dosage, monitoring, classroom routines and how learner data are acted upon, rather than simply through the presence of a teacher or a one-off training event.

Additional analyses address teacher influence by reporting that paraprofessionals could be as effective as teachers in some tutoring contexts; that different teacher-selected modes of usage affected **learner performance**; that classroom assistance varied; and that some studies either controlled for teacher fixed effects or minimised teacher presence to separate intervention effects from teacher effects (Wang et al., 2024; Huang, 2025). Overall, the evidence suggests that teacher involvement matters most when it shapes how ITSs are embedded into instruction, how learner data are acted upon, and whether pupils engage with the system consistently over time.

Parents' influence on impact was **rarely examined** beyond description; the explicit parent-related associations made are that marital status affected academic outcomes, and that parents with more education were more likely to help with distance education. The remaining parent-related findings concern consent, acceptability, feedback or home support rather than tested impact pathways (Ezzaim et al., 2023; Meeter, 2021). For this reason, the reviewed evidence does not support strong conclusions regarding parents as moderators of the effectiveness of ITSs. Their role appears more limited and context-dependent, mainly concerning consent, access, home support and acceptance of the technology.

In summary, the evidence suggests that teachers are the central adult actors in implementation and can influence outcomes under certain conditions. Teacher training and professional support therefore emerge as relevant implementation conditions, although the studies rarely isolate training effects causally. The strongest conclusion is therefore not that teacher training independently improves

ITS outcomes, but that teacher preparation, ongoing support and organisational capacity help to determine whether ITSs are used regularly, interpreted pedagogically and integrated into classroom routines. In contrast, parental involvement was described far more often than it was empirically examined as a factor shaping impact.

5. Discussion

This chapter of the report discusses and interprets the findings of the systematic review, focusing on key results, their conceptual integration, and their implications for understanding the impact of intelligent tutoring systems (ITSs).

5.1. Summary of key findings

Within the scope of this systematic literature review, the material reviewed provides a **substantial and strongly intervention-oriented evidence base** on ITSs in school education – although this evidence base is unevenly distributed and should be interpreted with some methodological caution. As set out in Section 1.3, the review does not provide primary research or a comprehensive account of all recent developments in AI-supported education, but instead focuses on ITSs for which causal evidence in school settings could be identified. It is dominated by quasi-experimental studies and randomised controlled trials, which together account for 93 % of the studies included. Therefore, the reviewed literature centres methodologically on comparative evaluations of instructional interventions, often using pre/post or baseline/endline measurement structures, rather than on purely descriptive or correlational research. Since quasi-experimental designs form the largest share of the studies reviewed, parts of the evidence base remain sensitive to selection bias, baseline imbalance and confounding.

Furthermore, the literature is **geographically concentrated**: the United States is the most prominent single country, while Asia is the most strongly represented region, with Thailand standing out as a major contributor. The evidence also focuses in particular on lower-secondary education, and is strongly weighted towards mathematics. Across outcome domains, the evidence is most extensive for learning outcomes, which are examined in almost all studies. Evidence on motivation and engagement is less consistently developed, and the evidence base on well-being and socio-emotional outcomes is clearly smaller. Lastly, the field appears to be expanding rapidly, with publications increasing markedly after 2020 and especially from 2024 to 2025 – although years in which data collection actually occurred are often not reported and should therefore be interpreted cautiously in light of the descriptive evidence mapping in Section 3.1 (subsections 3.1.1–3.1.6) and the limitations discussed in Section 4.3.

A second key finding focuses on the conceptual status of ITSs in the reviewed literature. The definitions of ITSs used by the studies share a common adaptive core, generally describing ITSs as computer-based or AI-based adaptive learning tools that personalise learning by responding to learner characteristics, performance, progress or estimated learner state. What distinguishes ITSs from broader AI-driven educational technologies is therefore not adaptivity alone, but

the combination of learner modelling, domain- or task-related structure, pedagogical decision-making, and responsive instructional support. At the same time, the reviewed studies show that these elements are operationalised in diverse ways, ranging from canonical model-based tutoring architectures to dialogue-based, diagnostic, platform-based, hybrid, affective or gamified systems. The present review therefore suggests that ITSs should be understood not as a single, clearly bounded type of intervention, but as a **heterogeneous family of learner-responsive instructional systems with a shared adaptive core** (Section 3.2). The implication of this for the report's synthesis is that the effects of these systems should not be attributed to "ITSs" as a uniform technology, but should be interpreted in relation to the specific adaptive, pedagogical and interface functions through which the systems support learning.

A third key finding focuses on the overall pattern of reported effects. Across the studies reviewed, findings are broadly favourable but are clearly differentiated by outcome domain. The strongest and most extensive evidence relates to **learning outcomes and skill development**, where ITSs are frequently associated with improvements in achievement, mastery, conceptual understanding, problem solving, reasoning, reading-related performance, computational thinking, digital literacy and other cognitive or process-oriented indicators. Where standardised effects are reported, these range from small to very large. However, the findings of the studies are not uniform: effects vary by the intensity of the intervention, implementation quality, comparison condition, learner subgroup, curricular unit, topic difficulty and specific outcome measure. Some studies report no significant between-group advantage, no added benefit of one intervention variant over another, or isolated weaker and negative effects. Therefore, this review supports neither a blanket claim of effectiveness nor a null position, but rather a **conditional pattern in which positive learning effects are common, yet are dependent on design, dosage, context and use** (subsection 3.3.1).

Motivation and engagement also show a generally positive pattern, but the evidence here is more heterogeneous in terms of both conceptualisation and reported results. The clearest positive motivational findings refer to enjoyment, hope, situational interest, self-efficacy, motivation, attention and confidence, which often show small-to-moderate effects in the most clearly reported studies. Behavioural engagement is also frequently reported as improving, including participation, time on task, practice intensity, progression, persistence, frequency of interaction, and continued use. However, this strand of evidence is less consistent than that for the learning domain, partly because the studies reviewed examine a wider and less uniform range of affective-motivational and behavioural indicators. Some motivational subscales remain unchanged in otherwise positive studies, and potential novelty effects are not consistently examined – especially in short, pilot, first-use or technology-rich settings. In addition, some engagement indicators vary by subgroup, implementation design, instructional condition or unit, with isolated negative or countervailing patterns such as gaming-the-system or errors of omission. The most appropriate summary is therefore that motivation

and engagement are often positively affected, but in a less stable and more context-sensitive way than learning outcomes (subsection 3.3.2).

Socio-emotional outcomes constitute the smallest cluster of evidence in this review. However, within this limited subset, the pattern reported is consistently more favourable than adverse or mixed. Where such outcomes were examined, studies report increases in positive affect, emotional well-being during learning, motivation, interest, confidence, self-efficacy, perceived competence, goal setting, favourable attitudes towards tutoring systems and interaction with affect-supporting agents. They also report reductions in negatively valenced or strain-related outcomes, including mathematics anxiety, frustration, effort, mental demand and cognitive load. These results should be interpreted cautiously, however, as they are based on a much smaller and less consistently measured evidence than the other outcome domains. Nevertheless, the available evidence does not suggest a predominantly adverse socio-emotional pattern, but instead points to a limited and broadly positive stream of evidence (subsection 3.3.3).

A fourth key finding is that the effects reported are best understood through the **mechanisms** by which ITSs translate learner information into instructional support. Across the studies reviewed, feedback, adaptive sequencing and scaffolding emerge as the most consistently recurring pedagogical mechanisms shaping learning processes. However, many different approaches are taken to the implementation of such mechanisms in ITSs, and not all of these approaches work equally well across all situations and for all users. This makes thoughtful and evidence-informed design essential. Some characteristics are identified in the reviewed literature as being particularly significant. Feedback appears particularly relevant when it is timely, specific, explanatory and connected to learners' errors, misconceptions or next steps. Adaptive sequencing matters where task order, difficulty, pacing or learning pathways are adjusted to learner performance, readiness, mastery or learner-model estimates. Scaffolding appears beneficial where support is appropriately tailored through hints, prompts, worked examples, guided practice, dialogic support or step-by-step assistance – although it is less often isolated as a distinct explanatory factor. Together, these mechanisms represent the most direct ways in which ITSs operationalise adaptivity and personalise instruction (subsections 3.4.2–3.4.4).

In contrast, the ways in which **technological design features** such as dashboards, progress displays, avatars, conversational agents, gamified elements, multimedia and augmented-reality interfaces appear to matter are a more conditional. The reviewed studies suggest that these features are most useful when they make adaptive support more visible, usable, engaging, or actionable – particularly when they reinforce feedback, mastery progression, self-monitoring, explanation or guided practice. In the studies, dashboards and conversational agents often supported regulation, engagement or affect, while gamification, multimedia and workflow-related features showed more selective effects within broader instructional systems. Since many studies evaluated complete bundles of features rather than isolated interface elements, no

conclusions can be drawn about the effectiveness on learning outcomes of individual features (subsection 3.4.5).

Lastly, the present review shows that the consistency of the effects of ITS use depends strongly on **implementation and context**. The evidence is clearest with regard to classroom-based use, while homework, informal and extracurricular uses are less common and less systematically examined. Teacher involvement emerges as a particularly important implementation condition: teachers frequently monitor progress, interpret dashboards, align use of the system with curricular goals, provide follow-up support, support the fidelity of the implementation, and shape how ITSs are integrated into instruction. Parental involvement, in contrast, is far less central in the evidence and is only rarely examined as an independent determinant of outcomes. The intensity of the implementation also appears as a relevant cross-cutting condition, since dosage, exposure, sustained participation and completion of practice repeatedly shape whether or not positive effects emerge, and how large these are. Educational stage and subject domain further structure the heterogeneity observed, with some studies reporting stronger effects in lower grades, as well as differences between subjects and variation across subdomains, task types, or learner groups. Taken together, these findings indicate that ITS mechanisms and outcomes are strongly conditioned by contextual, implementation-related and moderating factors (subsections 3.4.1, 3.4.6–3.4.8).

Overall, the key findings of the review point to a broadly favourable but clearly conditional pattern of evidence. ITSs are best understood as heterogeneous adaptive instructional systems, the effects of which depend on how learner information is translated into pedagogical support.

5.2. Synthesised impact framework

The conceptual framework introduced in subsection 1.1.2 distinguished between underlying mechanisms, observed effects and moderating or contextual factors.

The findings of this review broadly support that structure, but they also allow these relationships to be specified. Table 1 (shown in subsection 1.2.1) presents the initial conceptual impact framework used to guide the review. This distinguishes, at a relatively general level, between underlying mechanisms, moderating and contextual factors, and observed effects, and thereby provides a heuristic structure for organising the empirical literature on ITSs.

Table 3 below builds on this preliminary framework by refining each of these dimensions in the light of the review's findings. It specifies the category of underlying mechanisms more precisely than in Table 1 by distinguishing a diagnostic–adaptive layer, core pedagogical mechanisms and an interaction layer. It also expands the moderating and contextual dimension by identifying the intensity of implementation as an additional cross-cutting condition, and it

differentiates the observed effects more clearly by indicating the relative strength and function of the three outcome domains. In this sense, Table 3 should be read not as a replacement of the framework introduced in Section 1.2, but as its empirically grounded elaboration and specification.

Across the studies reviewed, this logic is most consistently visible in a functional configuration that is built around learner modelling, pedagogical or tutor control, and communication or interface functions. In a substantial part of the literature, a domain or expert model is also explicitly articulated; in other cases, this remains conceptually present but is less clearly specified. This means that the educational impact of ITSs is not adequately captured by the labelling of the system alone. What matters more is whether – and if so, how – learner information is translated into adaptive instructional support.

Table 3: Refined impact framework for ITSs

Dimension	Elements	Role in the impact logic
Underlying mechanisms	Diagnostic–adaptive layer: Learner modelling Performance monitoring Estimation of state Core pedagogical mechanisms: Feedback (timing, specificity, explanatory quality) Adaptive sequencing (task order, difficulty, pacing, pathways) Scaffolding (hints, prompts, worked examples, step-by-step support) Interaction layer: Technological design features (dashboards, conversational agents, gamification, multimedia interfaces)	ITSs translate learner data into adaptive instructional support. The diagnostic–adaptive layer enables personalisation. Core pedagogical mechanisms directly shape learning processes. Design features influence usability, visibility and engagement, as well as mediating interaction with the system
Moderating and contextual factors	Educational stage and age group Subject domain Educational context (classroom, homework, extracurricular) Teacher role and parental involvement – implementation intensity (dosage, exposure, sustained use)	These factors condition whether and how mechanisms translate into outcomes. They influence implementation quality, learner responsiveness and system use. Implementation intensity emerges as a key cross-cutting condition for effectiveness
Observed effects	Learning outcomes/effectiveness (achievement, skills, mastery, efficiency) Motivation and engagement (affect, persistence, time on task, participation, dropout) Well-being / socio-emotional outcomes (confidence, affect, anxiety, cognitive load)	Outcomes emerge as the result of how effectively the mechanisms align instruction with learners' needs. The strongest and most consistent effects are seen on learning outcomes. Motivation and engagement function both as outcomes and as mediating processes. Socio-emotional outcomes are less frequently studied but are predominantly positive

Source: Authors' own analysis.

The **core mechanisms** are, above all, adaptive sequencing, feedback and scaffolding. Adaptive sequencing refers to the adjustment of task order, difficulty, pacing or pathway in response to the learner's state. Feedback refers to the timing, specificity and explanatory quality of the system's responses to learner input. Scaffolding includes hints, prompts, worked examples, step-by-step assistance and other forms of structured support. Together, these mechanisms explain how diagnosis is converted into pedagogical action.

Technological design features form part of the broader mechanism structure, but they do not operate in the same way as the core pedagogical mechanisms. Dashboards, conversational agents, avatars, gamified elements, multimedia environments and other visible design features mainly shape visibility, usability, social presence and willingness to engage. They are best understood as an interaction layer that supports adaptive instruction when tied to feedback, mastery progression, self-monitoring or guided practice.

On this basis, the synthesised framework suggests that ITSs influence pupils' outcomes through a sequence of connected processes rather than through a direct technology-to-outcome relationship. At the centre of this sequence lies a diagnostic and adaptive layer. ITSs monitor learner performance, progress, errors and, in some cases, broader behavioural or affective indicators. These inputs are used to estimate learner state and to determine what kind of support should be provided next. In empirical terms, this **diagnostic-adaptive layer** forms the enabling condition for the core pedagogical mechanisms identified throughout the review.

This translation of learner information into adaptive instructional support is also shaped by **educational context**. Most of the available evidence refers to classroom-based use, sometimes supplemented by school-based tutoring or structured additional sessions. While homework, informal learning and extracurricular applications are present in the literature, they are less common and less systematically examined as distinct implementation settings. Therefore, the evidence base speaks most clearly about ITSs as part of formal instructional arrangements rather than as context-independent learning technologies.

Teacher involvement emerges as a particularly important condition for successful implementation. The reviewed studies suggest that ITSs do not simply replace pedagogical judgement; rather, they redistribute instructional work. Teachers often mediate system use through monitoring, the interpretation of dashboards, alignment with curricular goals, follow-up support and implementation management. Where teacher buy-in, fidelity and active orchestration are present, outcomes tend to be more favourable. Where time, infrastructure, competence or commitment are limited, effects appear more constrained. Parental involvement, in contrast, is far less central in the reviewed evidence, and is only rarely examined as an independent determinant of outcomes. It appears mainly as a procedural or supportive background condition rather than as a systematically tested driver of impact.

A further cross-cutting condition that is less visible in the initial conceptual framework but becomes prominent in the review findings, is implementation intensity. Across several studies, dosage, exposure, sustained participation and completion of practice appear to shape whether or not positive effects emerge, and how large these are. This suggests that the effectiveness of ITSs depends not only on the presence of adaptive functionality, but also on whether learners encounter enough well-aligned opportunities to benefit from it.

Taken together, these findings suggest that both mechanisms and outcomes are strongly conditioned by contextual and moderating factors. Educational stage and age group matter because adaptive support interacts with learners' developmental readiness, prior knowledge and self-regulatory capacity. The reviewed evidence suggests somewhat stronger or more visible effects in primary and lower secondary education than in later grades, although this pattern is not uniform across all studies. Subject domain also matters: the concentration of evidence in mathematics and other structured domains suggests that ITSs may be especially compatible with learning environments in which tasks can be decomposed, sequenced and evaluated against relatively explicit criteria. Positive findings in reading, language learning, science and programming indicate that the approach is not confined to one domain, but the mechanisms through which it operates may vary along with the structure of a given discipline.

This process-oriented view helps to explain the pattern of effects observed. The strongest and most extensive evidence refers to **learning outcomes and skill development**, whereby ITSs are often associated with improved achievement, conceptual understanding, problem solving, mastery or other cognitive indicators. **These effects remain conditional on learner group, implementation quality, dosage, task structure and comparison condition.**

Motivation and engagement form a second, related outcome domain. Even here, the overall tendency is positive, but the evidence is more heterogeneous. This is partly due to the diversity of constructs included under this heading. These range from enjoyment, self-efficacy and interest to persistence, time on task, participation and continued use. The findings suggest that ITSs are especially able to support these outcomes when they provide timely feedback, visible progress, learner choice or interactive forms of support. At the same time, motivational and engagement outcomes appear to be particularly sensitive to design choices and implementation conditions. Therefore, they may function both as outcomes in their own right and as proximal processes that help to sustain learning gains over time.

Socio-emotional outcomes constitute the smallest cluster of evidence in this review. Nevertheless, the available studies more often point to positive than to mixed effects, with reductions reported in anxiety, frustration, or cognitive load. However, this evidence should be interpreted cautiously because it is limited in scope and sometimes overlaps with motivational or engagement-related measures. The main implication is therefore not that ITSs automatically improve pupils' socio-emotional development, but that well-designed support may shape how supported and emotionally secure learners feel during learning.

Overall, the review supports a synthesised impact framework in which ITSs are understood as adaptive instructional systems that generate effects through a combination of learner modelling, pedagogical decision-making and responsive interaction. The evidence suggests that ITSs should not be understood as a single, uniform type of intervention. Rather, they are best conceptualised as a

heterogeneous family of adaptive instructional architectures which vary in format, interface and computational approach, while sharing a common instructional logic: they use information about the learner to shape pedagogical responses.

The most consistent contribution made by ITSs lies in improving the fit between learners' needs and instructional support. Where this fit is achieved through well-calibrated sequencing, timely feedback, appropriate scaffolding and purposeful teacher mediation, positive effects are most likely to emerge. Where the aforementioned conditions are weaker, effects are more mixed, more localised, or more dependent on specific outcome measures. The synthesised framework thus refines the initial conceptual model by making the diagnostic layer more explicit, distinguishing more clearly between pedagogical mechanisms and interface-related interaction features, and identifying the intensity of implementation as a cross-cutting condition for effectiveness.

5.3. Limitations of this review

Several limitations should be taken into account when interpreting the findings of this review.

First, the search strategy relied primarily on Elicit as its scientific search tool and, by extension, on the Semantic Scholar database. Although this approach enabled the highly structured and scalable identification of empirical studies, it may have limited the comprehensiveness of the evidence base. Studies indexed in other major databases but not adequately captured in Semantic Scholar may have been missed.

Second, the review applied deliberate eligibility restrictions with regard to publication period, language and population. By only including studies published in English between 2020 and 2026, the review may have excluded relevant evidence published in other languages or prior to the selected timeframe. Although these restrictions were appropriate for ensuring topical relevance and feasibility, they inevitably limit the international and historical width of the evidence base. Similarly, the review's exclusive focus on learners in primary, lower-secondary and upper-secondary education means that its findings cannot readily be generalised to early childhood education, higher education, vocational education or adult learning contexts.

Third, the screening procedure combined automated and manual elements, which introduces both efficiencies and potential sources of bias. The initial screening of 9,985 abstracts was conducted through automated classification in Elicit, with each record being assigned a relevance score and only the top 400 publications proceeding to manual abstract review. While this procedure substantially increased the feasibility of the review, it also creates the possibility that some relevant studies were not prioritised highly enough by the automated system and were therefore excluded from closer inspection. With this in mind, the

review is not only dependent on the predefined screening criteria, but also on the performance and assumptions made by the AI-supported ranking process. This should be borne in mind when considering the completeness of the final study set.

Fourth, although manual abstract screening and full-text screening were conducted by two reviewers, the review remains subject to reviewer judgement in its interpretation of the inclusion criteria, particularly with respect to the operationalisation of ITSs. The distinction between ITSs, AI-based adaptive learning tools and other forms of AI-supported educational technology is not always clear-cut in the reviewed literature. As a result, inclusion and exclusion decisions may have involved interpretive judgement, especially in borderline cases. This issue is amplified by the exclusion of systems based on generative AI models, which was justified methodologically as a means of maintaining conceptual clarity, but may limit the relevance of the review's findings to the rapidly evolving contemporary landscape of AI-supported learning technologies – particularly as generative AI may increasingly be integrated into future ITS architectures.

Fifth, a further aspect to consider is the distribution of evidence across outcome domains and moderators. While learning-related outcomes are reported by a comparatively large number of studies, socio-emotional outcomes and some subject domains (e.g. computer science) are represented to a lesser extent. As a result, conclusions in these areas are based on a more limited evidence base and should be interpreted with appropriate caution.

Sixth, a further limitation concerns the interpretation of survey-based, self-report, correlational and cross-sectional components within the evidence base. Although these components were included where they formed part of broader empirical evaluations, they do not provide the same basis for causal inference as experimental or quasi-experimental comparisons. This is particularly relevant with regard to motivation, engagement, attitudes, learner experience and intention to continue using the service, which were often measured through questionnaires, self-reports, modelled associations or cross-sectional assessment structures. Therefore, the synthesis treats such findings as complementary contextual evidence.

Seventh, the interpretation of reported effects is further limited by the extent to which the primary studies included account for novelty effects. Several studies involved short interventions, first-time use, pilot settings or technology-rich learning environments, while potential novelty effects were often not explicitly examined or controlled for. This is especially relevant with regard to motivation- and engagement-related outcomes, where positive responses may partly reflect pupils' initial interest in a new or interactive technology rather than being sustained effects of the instructional design. At the same time, studies involving repeated use or longer intervention periods provide a more suitable basis for examining whether or not effects persist beyond initial exposure, although they

do not necessarily isolate novelty effects directly. As a result, this review can identify novelty as a possible interpretive limitation, but it cannot systematically assess the magnitude or direction of novelty effects across the included evidence base.

Eighth, the review is constrained in the conclusions it can draw about sustained adoption and routine use. Because the research question prioritised causal evidence, the studies included are mainly pilots, trials or structured interventions rather than evaluations of long-term everyday use. Accordingly, the review can assess only to a limited extent whether ITSs are adopted, maintained and integrated into school practice beyond research settings. Issues such as curricular integration, teacher capacity, infrastructure, implementation support and sustained use are already visible in the reviewed evidence as important contextual conditions. Other potential barriers to adoption, including development and maintenance costs or ethical and privacy concerns, are less directly captured by the included studies, and therefore remain beyond the scope of this review's systematic assessment.

Ninth, the review did not use AI tools merely for search and screening, but also for data extraction and evidence synthesis. The extraction matrix was applied in Elicit, and the synthesis of extracted content was supported by GPT-5.4 through processes of summarisation clustering, and transformation into academic prose. Although these steps were used to support reviewer-led analysis rather than replace it, AI-assisted extraction and synthesis may introduce omissions. To reduce this risk, the AI-supported steps were accompanied throughout by human oversight, critical verification and interpretive judgement. Therefore, the validity of the resulting synthesis rests not on automated processing alone, but on the extent to which these stages were actively reviewed and checked by the researcher.

Lastly, the present review prioritised narrative and thematic synthesis over formal quantitative aggregation. This is appropriate given the diversity of interventions and outcomes, but it also means that the review cannot provide pooled effect estimates or statistically test the magnitude of effects across studies. Consequently, the conclusions should be understood primarily as a structured interpretive synthesis of the available evidence rather than as a meta-analytic estimate of effectiveness.

Taken together, the above limitations do not invalidate the review's findings, but they do define the boundaries within which its conclusions should be interpreted. The review provides a structured and methodologically transparent synthesis of a substantial body of recent empirical research, but its conclusions should be read in light of the restrictions on database accessibility, the use of AI-assisted review procedures, the conceptual boundaries applied to ITSs, the limited consideration of novelty effects in the primary studies, and the heterogeneity of the underlying evidence base.

5.4. Conclusions

Taken together, the causal evidence indicates positive effects more often than null or negative effects, particularly with regard to learning outcomes and skill development. However, the direction, magnitude and consistency of these effects vary across outcome domains, system designs, learner groups, subject areas, levels of use and implementation contexts. ITSs should therefore not be treated as a single, stable intervention category, but as a heterogeneous family of adaptive instructional systems that use information about the learner to shape pedagogical responses.

Against this background, an answer to the research question can be derived from the reviewed literature. The existing causal empirical evidence on general education identifies predominantly positive effects of ITSs on pupils' learning outcomes, generally positive but more heterogeneous effects on motivation and engagement, and a smaller but largely favourable body of evidence on socio-emotional development.

The strongest and most extensive evidence refers to educational outcomes and skill development. Across the studies reviewed, ITSs were most often associated with improvements in achievement, conceptual understanding, problem solving, mastery, reasoning, reading-related performance, computational thinking, scientific competency and other cognitive indicators. Positive effects were found in mathematics as well as reading, language learning, science and computer science-related contexts, although the evidence outside mathematics is thinner and should be interpreted with caution. Effects varied by learner group, intervention intensity, comparison condition, curricular unit, topic difficulty and outcome measure. Therefore, the overall conclusion here is not that ITSs consistently outperform all alternatives, but that positive effects on learning are common when systems are well aligned with learners' needs and are implemented with sufficient intensity.

The evidence on motivation and engagement is also generally positive, but is more heterogeneous in both conceptualisation and results. ITS-supported conditions were often associated with higher enjoyment, situational interest, self-efficacy, confidence, motivation, attention, participation, time on task, progression and continued use. Socio-emotional and well-being-related outcomes were examined far less often, but the available findings more often point to positive affective outcomes and reductions in anxiety, frustration, effort, mental demand and cognitive load. These findings should be interpreted cautiously because this evidence base is smaller and less consistently measured.

The review further indicates that effects are best understood through the mechanisms via which ITSs operate rather than through the mere presence of an AI-based or "intelligent" system. Across the literature, the most consistent functional pattern is built around learner modelling, pedagogical control, and

communication or interface functions. Adaptive sequencing, feedback and scaffolding emerge as the core mechanisms most closely linked to positive outcomes. Feedback appears especially important when it is timely, specific, explanatory and connected to learners' errors or next steps; adaptive sequencing matters when task order, difficulty, pacing or pathways are tailored to the learner's state; and scaffolding appears beneficial when support is graduated and well calibrated.

A key conclusion of the review is that variations in effect depend to a large extent on implementation and context. Most studies examined classroom-based use, whereas homework, informal learning and extracurricular applications were less common and less systematically studied. Teacher involvement emerged as a particularly important implementation condition: teachers monitored progress, interpreted dashboards, aligned system use with curricular goals, provided follow-up support and shaped the fidelity of implementation. While teacher training and professional support appear relevant to the quality of implementation, these were rarely isolated as independent causal factors. Implementation intensity is another cross-cutting factor, given that dosage, sustained participation, exposure, and completion of practice shaped whether benefits emerged and how large they were. Effects also varied by educational stage, subject domain and learner subgroup. Learners' digital skills are another relevant boundary condition, but these were rarely examined systematically in the studies reviewed.

A related structural limitation applies to the translation of advances in research into sustained and scalable educational practice. Although the reviewed literature documents progress in adaptive feedback, sequencing, scaffolding, learner modelling and interface design, many ITSs remain research prototypes, context-specific platforms or systems evaluated under structured intervention conditions. Positive findings from causal studies therefore cannot automatically be interpreted as evidence that ITSs will generate similar effects under routine classroom conditions.

The overall conclusion, then, is a cautious but clearly positive one. The causal evidence reviewed supports the conclusion that ITSs can contribute meaningfully to pupils' learning outcomes in general education and can also support motivation and engagement and, in a smaller body of studies, their socio-emotional development. Yet the same evidence also shows that these benefits are not automatic consequences of digitalisation or technical novelty. The most consistent contribution made by ITSs lies in improving the fit between learner needs and instructional support, and this contribution is most visible when adaptive functionality is translated into well-calibrated sequencing, timely and specific feedback, appropriate scaffolding, and purposeful pedagogical mediation. Where these conditions are weaker, effects are more mixed, more localised, or are more dependent on the outcome measure and comparison condition concerned. This review therefore justifies neither a blanket endorsement of ITSs nor blanket scepticism, but rather a conditional conclusion:

that the literature supports the educational potential of ITSs, while also showing that this potential is realised unevenly and under identifiable pedagogical and implementation conditions. Therefore, the most robust takeaway is not that ITSs are effective *per se*, but that they are effective insofar as they achieve a high degree of alignment between learner modelling, pedagogical response and instructional context.

At the same time, the conclusions of the review must be bounded by the structure of the evidence base itself. The literature predominantly focuses on lower-secondary education, mathematics and studies conducted in Asia and the United States, while Europe and several other regions are less strongly represented. Most evidence also concerns structured, school-based interventions rather than routine use across diverse institutional contexts. Therefore, the findings of the review should be interpreted as being evidence-informed rather than universally transferable in a simple one-to-one manner.

The above conclusions apply most directly to individually oriented ITSs, which dominate the evidence base reviewed. Cooperative or collaborative ITSs constitute a related but distinct paradigm because they extend adaptive support from individual learner–system interaction to peer interaction, shared problem solving and group-level knowledge construction. By way of contextual evidence that goes beyond the systematic review corpus, recent work shows that collaborative ITSs may allow learners to act both as learners and as sources of explanation for peers, supporting social knowledge building through the discussion and negotiation of meaning, and fostering communication, critical thinking, problem-solving skills, motivation and teamwork (Arnau-González et al., 2025; Haq et al., 2020).

Such systems also require additional design mechanisms that go beyond individual learner modelling. These include learner–learner interaction, tutor–group interaction, shared solution spaces, peer commenting or rating, real-time synchronisation and the orchestration of collaborative activity (Arnau-González et al., 2025; Haq et al., 2020). Future research should examine more explicitly how cooperative ITSs relate to the mechanisms identified here and whether or not they generate distinct learning, motivational or socio-emotional effects in school-based settings.

A further future-facing issue relates to the integration of generative AI into ITSs. Generative AI systems were excluded from this review to maintain conceptual clarity, but recent literature suggests that they may increasingly reshape ITSs by supporting more natural dialogue, dynamic explanations, personalised feedback, richer learner-state inferences, and multimodal or affect-sensitive forms of support. As contextual evidence beyond the systematic review corpus, Batsaikhan and Correia (2024) and Maity and Deroy (2024) indicate that these developments may extend, rather than replace, the core mechanisms identified in this review – in particular, feedback, scaffolding, sequencing and learner modelling. Future research should therefore examine if generative or hybrid ITS

architectures can be designed and used in pedagogically meaningful and ethically governed ways that enable robust benefits under ordinary school conditions.

6. Implications and recommendations

The review provides a substantial but bounded evidence base for policy decisions about ITSs, including related AI-based adaptive learning tools in general education. Building on the synthesised impact framework, this chapter derives implications and recommendations for policy and practice, focusing on how mechanisms, implementation conditions, and contextual factors shape the effective use of ITSs in school education.

The recommendations that follow distinguish between three levels of responsibility. National and regional policymakers – including ministries, regional education authorities and relevant public funding, procurement, quality-assurance and professional-development bodies – shape funding frameworks, procurement and accountability requirements, evaluation expectations and capacity building. School authorities and school leaders make or implement local selection decisions, integrate the use of ITSs into curriculum planning and timetables, assign implementation roles, support teacher orchestration, and monitor use and outcomes. European Union institutions can add value by supporting common reporting principles, comparable and multi-country evidence, reusable evaluation resources, and shared evidence infrastructure, while respecting the competence of Member States over curricula and implementation. The allocation of these responsibilities varies between education systems; thus, the categories used in this chapter refer to areas of responsibility rather than to uniform institutional structures.

Throughout this chapter, recommendations that follow directly from the reviewed evidence are linked to the relevant sections of the synthesis. Recommendations on conditional scaling, shared evidence infrastructure and legal or governance safeguards are evidence-informed policy judgements: they respond to limitations or implementation risks that were identified in the review but were not themselves evaluated as interventions in the included studies (Sections 3.1 and 4.3).

6.1. Policy-relevant takeaways

The review findings point to five policy-relevant areas: the intended use and observed effects of ITSs; the pedagogical mechanisms through which ITSs operate; the implementation conditions needed for the use of ITSs; the evaluation of the effects of ITSs under the intended conditions; and the transferability of the review's findings to European education systems. These areas are analytically related and should be considered together (Section 4.2).

6.1.1. Specify the intended use and observed effects before selecting an ITS

The evidence does not support a general conclusion that ITSs are equally effective across purposes, subjects, learner groups, curricular units or comparison conditions. Evidence is strongest in mathematics and with regard to learning outcomes, while findings in other subject areas and with regard to motivation, engagement and socio-emotional outcomes are less extensive or less consistent (subsections 3.1.5–3.1.6, 3.3.1–3.3.3 and 3.4.7). Before considering a particular system, national and regional authorities should define the educational or instructional need to be addressed and explain why an ITS is expected to add value in relation to realistic alternatives. This need may, for example, relate to insufficient opportunities for individualised practice feedback, limited capacity to differentiate particular practice phases, or a specified competency gap. National and regional authorities should therefore require the intended use to be specified prior to funding, procuring, or approving an ITS. This intended use should identify the pupil group, subject domain, educational setting, outcome domain, comparison condition and outcome measures.

This requirement could be implemented through funding criteria, procurement templates, digital-education strategies, and approval or quality-assurance processes. The comparison condition should represent the realistic alternative available in the intended setting, such as existing teacher-led practice, another digital tool, or current support arrangements. This does not necessarily require a new controlled trial for every procurement; it requires authorities to state what the ITS is expected to improve compared with current practice, and how that added value will be assessed. Progress should therefore be assessed against the specified outcome and the expected added value relative to this alternative, rather than against system use or user satisfaction alone. Authorities should also require claims to be limited to the subject area, learner group and outcome for which supporting evidence exists.

6.1.2. Translate pedagogical mechanisms into quality and procurement requirements

Across the review, educational value depends less on an “AI” or “intelligent” label than on how learner information is translated into instructional support. Timely and specific feedback linked to errors or next steps, adaptive sequencing based on performance or estimated mastery, and scaffolding through hints, prompts, worked examples, guided practice, dialogic support or step-by-step assistance recur across more favourable findings (subsections 3.4.2–3.4.5). Such support should respond to pupils’ current performance and difficulties, provide guidance in graduated steps, and be reduced as pupils become more competent. Rather than simply providing solutions, it should support active engagement with the task and, where appropriate, prompt pupils to work through or explain their reasoning.

By contrast, dashboards, avatars, conversational agents, gamification and multimedia do not show an independent and uniformly positive effect; they appear most useful when they make core instructional mechanisms more visible, usable or actionable.

National and regional authorities, public procurement bodies and funding agencies should therefore establish quality requirements that ask providers to explain how learner information is collected and interpreted, how feedback, sequencing, and scaffolding operate, how the system aligns with curriculum goals, what teachers can see and control, and what evidence supports the proposed use. With respect to scaffolding, providers should specify whether the system responds to pupils' errors or misconceptions, offers graduated hints before providing a solution, reduces support when performance or competence improves, and prompts pupils to elaborate or revise their responses rather than accepting superficial answers. Procurement and approval processes should also examine whether interface features strengthen the instructional function or merely add technical appeal. Monitoring should assess whether pupils receive the intended feedback and support for progression, whether teachers can act on system information, and whether these functions translate into the specified learning outcome. Since individual mechanisms were often evaluated as part of broader system packages rather than being isolated experimentally, these requirements should be treated as criteria for assessing pedagogical design, not as a guarantee of effectiveness.

6.1.3. Fund implementation capacity as part of the intervention

In the studies reviewed, access to software did not equate to effective implementation (subsections 3.3.1, 3.4.1, and 3.4.8). Greater exposure, sustained participation and completion of practice were repeatedly associated with stronger outcomes, whereas low-dosage or weakly integrated use produced smaller, mixed or null effects. Hybrid models combining human tutoring with ITS support produced stronger diagnostic gains than software-only use in some studies, and additional tutoring hours, improved tutor-to-pupil ratios or dedicated in-school supervision were associated with stronger use and outcomes. These findings indicate that implementation arrangements are part of the intervention rather than peripheral delivery details. They also indicate the importance of embedding ITSs within regular instructional practices and supporting their use over time, although the review does not establish a single implementation model or duration that is effective across settings. The evidence also speaks most clearly about classroom-based and school-embedded use. Proposals for homework, informal or extracurricular use should therefore be evaluated in their own implementation context rather than being assumed to reproduce classroom findings.

Funding and implementation plans should therefore include teacher preparation time, professional development, ongoing pedagogical and technical support,

scheduling, infrastructure, supervision and clearly assigned implementation roles. Support should extend beyond initial access to the technology or one-off training and should help schools connect ITS use with curriculum goals, timetables, classroom routines and teachers' instructional practices over time. Depending on national arrangements, this may include cooperation between schools, universities or other teacher-education institutions, and organisations responsible for school-based professional learning. The evidence is consistent with implementation models in which teachers retain pedagogical oversight and use ITSs as instructional supports; however, it does not establish teacher mediation as an independent causal factor. Authorities should require an intended pattern of use and should resource and monitor actual exposure, participation, completion, implementation fidelity, technical interruptions and teacher workload. The evidence does not support a universal minimum dosage; requirements should instead specify and test a plausible level of use for the intended purpose. Teacher training was rarely isolated in the studies as an independent causal factor. The review therefore does not establish a universally effective intensity or format for professional development, but neither does it support presenting one-off training as sufficient where organisational capacity, time, infrastructure or ongoing support are weak.

6.1.4. Pilot under routine conditions and make scaling conditional on evidence

Most of the studies included evaluated pilots, trials, prototypes or structured interventions rather than long-term routine use (Section 4.3). Reported effect sizes therefore provide evidence of the potential of ITSs, but are not a sufficient basis for broader transfer or scaling. National and regional authorities should begin with a limited pilot, evaluate the system under conditions comparable to its intended use, adapt the implementation where needed, and only then consider broader deployment. Before the pilot begins, authorities should specify the intended learning outcome and outcome measures, the comparison condition, the intended pattern of use, and the basis on which implementation will be continued, adapted or not extended. These criteria should reflect the factors identified in the review, including changes in the specified learning outcome, actual exposure and completion, teacher workload, infrastructure and support requirements, subgroup outcomes and unintended consequences. The review does not provide universal numerical thresholds for these criteria; they should be defined in relation to the intended use and implementation context. Evaluations should also distinguish initial motivation or engagement from sustained educational value – in particular, because potential novelty effects were rarely examined in short, pilot or first-use settings (subsection 3.3.2).

Publicly funded pilots should specify the comparison condition and should report not only whether effects are statistically significant, but also whether they are educationally meaningful for the specified learning outcome, together with effect sizes, participation and exposure, curriculum alignment, teacher workload,

infrastructure and support requirements, durability, subgroup outcomes and unintended consequences. Educationally meaningful effects are therefore changes that are relevant to the intended educational purpose, rather than system use, engagement or user satisfaction alone. This review does not establish a universal minimum effect size. Broader implementation should be conditional on evidence that the observed effects can be reproduced without exceptional project support and without widening differences in access, use or outcomes. Where the intended effects are not demonstrated, authorities should adapt the implementation model, revise the requirements, or discontinue the intervention.

6.1.5. Build a shared European evidence and coordination framework

European evidence is less extensive than evidence from the United States and selected Asian countries, and studies use heterogeneous outcomes, comparison conditions and implementation measures (subsections 3.1.2, 3.3.1–3.3.3 and 4.3). This limits the ability of national and regional authorities to compare systems, to learn from implementation experience, and to judge transferability. European Union institutions, working with Member States and relevant education authorities, should support common evaluation and reporting principles, comparative studies under routine European conditions, shared implementation indicators, and structured exchange on successful and unsuccessful adoption. Across these settings, the evidence that is most transferable is that which focuses on recurring pedagogical mechanisms and implementation conditions rather than particular products or platforms.

European coordination should focus on functions that add value without prescribing national curricula, centralising procurement decisions or implying that a single European model for ITS implementation is appropriate. These functions include guidance on evidence claims; common definitions of learning, engagement, implementation and equity indicators; reusable evaluation resources; transparent reporting templates; and repositories documenting implementation models as well as both successful and unsuccessful results. Equity, accessibility, privacy, transparency and the responsible use of data should be incorporated as cross-cutting safeguards within these activities. The present review does not support the assumption that personalisation automatically improves equity: although some lower-achieving pupils and disadvantaged groups have benefitted strongly from this, it was factors such as access, attendance, infrastructure, prior achievement and, potentially, learners' digital skills that determined who had sufficient exposure to an ITS and benefitted from it. European, national and regional authorities should therefore monitor access, sustained use, learning outcomes and unintended differences across learner groups. Progress at European level should be assessed through the comparing and reuse of evidence, the inclusion of routine-use and unsuccessful

cases, and the extent to which national and regional decisions can draw upon relevant European experience.

6.2. Recommendations for national and regional policymakers

Building on this report's policy-relevant takeaways, this section translates the findings of the review into recommendations for national and regional policymakers. While Section 5.1 identifies the main implications regarding intended use and observed effects, pedagogical mechanisms, implementation conditions, evaluation and transferability, this section sets out how these considerations can inform strategic alignment, funding and procurement, implementation capacity, professional development, equity and monitoring. Each subsection identifies the relevant evidential basis and distinguishes the direct implications of the review from broader evidence-informed policy judgements.

6.2.1. Intended use and strategic alignment

- Specify the learner need or curricular goal to be addressed, together with the intended uses, pupil groups, subject domains, educational settings, outcome domains and comparison conditions prior to allocating public funding or including ITS in digital-education strategies.
- Require funding applicants and procuring authorities to state why an ITS is expected to provide added value relative to the specified comparison condition, the instructional setting, curriculum link and outcome indicators for each proposed use.
- Limit public claims and approval decisions to the learner groups, subjects and outcomes for which relevant evidence is available.

The evidential basis for these recommendations lies in the uneven distribution of evidence across school stages, subjects and outcome domains (subsections 3.1.4–3.1.6 and 3.3.1–3.3.3), together with the variation identified by educational stage and subject domain (subsections 3.4.6–3.4.7).

6.2.2. Funding, procurement and quality requirements

- Establish procurement and quality criteria that cover learner modelling, domain or task structure, feedback, sequencing, scaffolding, curriculum alignment, teacher control, usability, accessibility and the evidence supporting the intended use. Require providers to explain how learner information changes the feedback, tasks or support that pupils receive, how these functions align with curricular goals, and what teachers can see and control.

- Require providers to report the comparison condition, study context, effect sizes (if available), intended pattern of use, implementation requirements, evidence limitations and known subgroup differences. Supporting evidence should be linked to the specified pupil group, subject domain, educational setting, outcome and actual level of exposure.
- Include the implementation conditions identified in the review in funding decisions. These should include preparation time, professional development, infrastructure, technical and pedagogical support, supervision and scheduling. Funding applications should specify how these conditions will be met, and who is responsible for them.
- Require providers to document whether and how teachers contributed to the pedagogical requirements, system testing and curriculum alignment, and what role teachers are expected to perform during implementation.

These recommendations are grounded in the synthesis of ITS components and features (subsections 3.2.2–3.2.3) and of feedback, adaptive sequencing, scaffolding and technological design (subsections 3.4.2–3.4.5). Direct comparisons support some feedback and sequencing mechanisms, whereas scaffolding and interface features have often been evaluated within wider system packages. The recommendations accordingly require mechanism-specific evidence without assuming that feature labels alone establish effectiveness. The translation of these findings into procurement and quality-assurance criteria is a policy application of the evidence rather than a tested procurement intervention.

6.2.3. Implementation capacity and professional development

- Fund teacher preparation and ongoing support as part of implementation rather than expecting schools to meet these requirements out of their existing resources.
- Define responsibilities for pedagogical leadership, technical support, session coordination, monitoring and escalation before the implementation begins.
- Integrate the pedagogical interpretation of dashboards and learner data into professional development, and test whether teachers can understand and act on system information.
- Monitor teacher workload and revise implementation models where technical, administrative, or monitoring demands displace core instructional work.

These recommendations are grounded in the implementation evidence synthesised in subsections 3.3.1, 3.4.1 and 3.4.8. As teacher training was rarely isolated as an independent causal factor, the recommendations on professional

development translate observed implementation requirements into policy practice rather than implying a separately demonstrated training effect.

6.2.4. Implementation, equity and monitoring

- Make wider implementation conditional on evidence from pilots conducted under conditions comparable to the intended use, and assess this evidence against criteria for continuation, adaptation or discontinuation specified before the pilot begins.
- Require evaluations to distinguish learning outcomes from motivation, engagement, satisfaction and usage, and to specify the comparison condition.
- Require reporting on actual exposure, completion, implementation fidelity, technical interruptions, curriculum alignment, teacher workload, infrastructure and support requirements, and the durability of effects.
- Monitor access, sustained use, outcomes and unintended consequences by relevant learner characteristics and school contexts; adapt or discontinue implementation where inequalities are widened.
- Ensure privacy, data protection, accessibility, transparency and safeguards against inappropriate profiling through the applicable legal, procurement and governance frameworks.

These recommendations draw on evidence concerning dosage, learner and contextual variation, and implementation conditions (subsections 3.3.1–3.3.2 and 3.4.1, 3.4.6–3.4.8), together with the limited evidence on sustained routine use (Section 4.3). The pilot-first approach is therefore an evidence-informed precautionary judgement rather than a scaling strategy directly tested in the studies reviewed. Privacy, data protection, accessibility, transparency and safeguards against inappropriate profiling are broader legal and governance requirements, and lie outside the causal evidence reviewed here.

6.3. Recommendations for schools in the European Union

This section translates the findings of the review into recommendations for school authorities and school leaders in the European Union. The evidential basis for these recommendations is primarily provided by subsections 3.3.1–3.3.3 and 3.4.1–3.4.8.

6.3.1. Intended use and system selection

- Before selecting an ITS, concretely define an intended pattern of use: the pupil group and subject domain; the curricular learning outcome to be supported; the instructional setting and activity in which the system will be used; the planned frequency or duration of use; how teachers will use system-generated information; and the comparison condition against which effects will be assessed.
- Assess systems on the basis of learner modelling, domain or task structure, feedback, sequencing, scaffolding, teacher control, usability and accessibility rather than an “AI” or “intelligent” label.
- Limit expectations to those learner groups, subject domains, settings, and outcomes for which relevant evidence is available.
- Test shortlisted systems with teachers and pupils, including the usability of dashboards, feedback and workflow arrangements, before committing to broader implementation.

The focus on intended use and system selection follows from the heterogeneity of effects across outcomes, educational stages, subjects and design configurations (subsections 3.3.1–3.3.3 and 3.4.5–3.4.7).

6.3.2. Teacher roles and implementation conditions

- Use ITSs as a teacher-mediated instructional support. Teachers should monitor learner progress, interpret dashboards and reports, connect system-generated information to curricular goals, and decide when follow-up instruction, additional support or direct intervention is required. System-generated information should inform rather than replace teachers’ pedagogical judgement.
- Assign responsibilities for pedagogical coordination, scheduling, supervision, reminders, technical troubleshooting, monitoring and maintaining implementation fidelity before implementation begins.
- Provide teachers with preparation time, professional learning and ongoing technical and pedagogical support. Professional learning should prepare teachers to interpret learner data, use dashboards, integrate the ITS into curriculum-aligned instruction, maintain classroom routines, and respond to technical or implementation problems.
- Integrate ITS use into curriculum planning and timetables, and define an intended pattern of use that provides pupils with sufficient and equitable opportunities to participate.
- Monitor teacher workload and adapt implementation where technical, administrative or monitoring demands displace core instructional work.

These recommendations are informed by evidence on teacher involvement, supervision, sustained use and organisational support (subsections 3.3.1, 3.4.1 and 3.4.8). While the studies do not generally isolate teacher involvement as an independent causal factor, the reviewed evidence indicates that it can shape how ITSs are embedded into instruction and how learner data are acted upon.

6.3.3. Pilot, monitoring and adaptation

- Begin with a limited pilot conducted under conditions comparable to the intended use before considering broader implementation.
- Distinguish learning outcomes from motivation, engagement, satisfaction and usage when assessing educational value.
- Monitor pupil exposure and completion, implementation fidelity, technical interruptions, curriculum alignment, teacher workload and differences between learner groups.
- Review implementation at predefined points and adapt, pause or discontinue use where the intended educational value is not demonstrated or if unintended inequalities emerge.

These recommendations draw on evidence that outcomes vary with patterns of use and implementation conditions (subsections 3.3.1–3.3.2), together with the limited evidence on sustained routine use (Section 4.3). The pilot-first approach is an evidence-informed precautionary judgement rather than an implementation strategy directly tested in the reviewed studies.

6.3.4. Transferability, equity and safeguards

- Do not assume that products, implementation models or reported effect sizes transfer directly between educational systems.
- Use the recurring pedagogical mechanisms and implementation conditions identified in the review as points of reference when adapting decisions to national curricula, school organisation, infrastructure, teacher roles and pupil characteristics.
- Monitor access, sustained participation, learning outcomes and unintended consequences across relevant learner groups and school contexts.
- Address accessibility, privacy, data protection, transparency and safeguards against inappropriate profiling through the applicable European and national legal and governance frameworks.

The transferability and equity recommendations draw on the geographical concentration of the evidence and on variation by context, educational stage,

subject domain, learner group, access and participation (subsections 3.1.2, 3.4.1 and 3.4.6–3.4.8). In contrast, privacy, data protection, accessibility, transparency and safeguards against inappropriate profiling are broader legal and governance requirements that were not systematically evaluated in the reviewed studies.

6.4. Recommendations for European Union institutions

While respecting Member States' competence over education, European Union institutions can add value through coordination, comparability and shared evidence. They should:

- Develop a common European reporting template for ITS evaluations, supported by shared definitions and minimum reporting fields for intended use, outcome measures, comparison conditions, effect sizes, implementation intensity, teacher workload, subgroup outcomes and evidential limitations.
- Create a shared European evidence and implementation repository documenting evaluation contexts, pedagogical mechanisms, patterns of use and implementation conditions, as well as both successful and unsuccessful results.
- Support comparative and multi-country evaluations of ITSs under routine European school conditions, including studies outside mathematics and lower-secondary education.
- Provide reusable evaluation guidance and sets of indicators that national and regional authorities can adapt to their own institutional and curricular contexts, together with evidence-informed guidance on funding and procurement based on pedagogical quality, implementation requirements, accessibility, transparency and responsible data use.
- Facilitate structured exchange between Member States and relevant education authorities on implementation models, barriers and adaptations, as well as successful and unsuccessful cases, without prescribing national curricula or centralising procurement decisions.

The development of evidence should support concrete policy decisions rather than operating as a general call for further research. Public funders and education authorities should:

- Commission multi-year evaluations of routine ITS use to examine whether or not effects persist after initial implementation and external project support have ended.

- Require evaluations to report total implementation and maintenance costs, teacher workload, infrastructure requirements, technical reliability and organisational capacity alongside pupil outcomes.
- Include teacher-facing outcomes that examine workload, the interpretation and use of dashboards, instructional decision-making and teachers' ability to act on system-generated information.
- Develop a common minimum set of outcome and implementation indicators that distinguishes learning, motivation, engagement, socio-emotional outcomes, exposure and the fidelity of implementation.
- Fund comparative studies that assess ITSs against relevant comparison conditions, including teacher-led instruction, other digital tools and different combinations of human and automated support.
- Require subgroup analyses that examine access, sustained participation, learning outcomes and unintended differences by learner starting point, educational stage, school context and relevant equity characteristics.
- Investigate the extent to which different levels and patterns of exposure affect educational outcomes, without assuming that there is a single "dosage" that applies equally to all systems, subjects or groups of learners.
- Document the failure, adaptation and discontinuation of implementations as systematically as successful adoptions, so that future policy decisions do not rely solely on positive cases.

These recommendations draw on the geographical concentration and heterogeneity of the evidence, the variation identified across learner groups, educational stages, subject domains and implementation contexts, and the limited evidence on sustained routine use (subsections 3.1.2, 3.3.1–3.3.3, 3.4.1, 3.4.6–3.4.8 and 4.3).

6.5. Implications for evaluating emerging generative AI-based tutoring

Systems based on generative AI- and LLMs were excluded from the operational definition of ITSs used in this review. Therefore, the findings provide no basis for conclusions regarding the effectiveness of generative AI-based tutoring. Contextual literature outside the systematic review corpus suggests that generative AI may extend structured ITSs through more flexible natural-language interaction, dynamic explanations, content variation and conversational support (Batsaikhan & Correia, 2024; Maity & Deroy, 2024). These represent potential capabilities rather than evidence of effectiveness in general education.

Nevertheless, the reviewed evidence on ITSs can provide an analytical framework for evaluating emerging generative AI-based tutoring. Rather than focusing on the generative or conversational interface itself, this framework directs attention to underlying mechanisms, moderating and contextual factors, and observed effects, drawing on the synthesis in subsections 3.4.1–3.4.8 and the impact framework developed in Section 4.2.

The central question with regard to the underlying mechanisms is whether learner information is translated into an appropriate pedagogical response – and if so, how. Evaluations should describe what information about learner performance, progress or errors is used; how generated responses are connected to a verified domain or task structure and curricular goals; and how this information shapes feedback, sequencing and scaffolding. Where instructional content is generated dynamically, evaluations should also report how its accuracy and pedagogical appropriateness are checked. Complementary research on LLM-based tutors supports this focus by distinguishing pedagogically appropriate guidance grounded in learner errors or confusion from plausible responses that function primarily as question-answering (Maurya et al., 2025).

Consistent with the preceding recommendations, evaluations should specify the intended use and the comparison condition, as well as documenting the relevant moderating and contextual factors. These include the role and control of teachers, actual exposure and completion, technical reliability, infrastructure and support requirements, teacher workload, and differences between learner groups. Evidence should be collected under conditions that are comparable to the educational settings in which the system is intended to be used.

Observed effects should be assessed separately from system capabilities and patterns of use. Emerging evidence synthesised by the OECD (2026) indicates that improved performance on a task with generative AI does not necessarily result in learning gains. Therefore, evaluations should distinguish learning outcomes from immediate task performance and from motivation, engagement, satisfaction and usage measures.

7. References

- Abdulayeva, A., Zhanatbekova, N., Andasbayev, Y., Khaimuldanov, Y., & Zhiyembayev, Z. (2025). 'The Role of Artificial Intelligence (AI) in Personalised Physics Education'. *Jurnal Pendidikan IPA Indonesia*, 14(3), 599-615 . <https://doi.org/10.15294/jpii.v14i3.29752>
- Alakayleh, S.S. (2025). 'The Impact of Artificial Intelligence-Based Tutoring Systems on Developing Mathematical Reasoning among Secondary School Students'. *American Journal of Applied Sciences*, 07(11), 85-99. <https://doi.org/10.37547/tajas/volume07issue11-10>
- Alenezi, A. (2024). 'Evaluating the Effectiveness of AI-Powered Adaptive Learning Systems in Secondary Schools'. *International Journal on Studies in Education*, 6(4), 686–700. <https://doi.org/10.46328/ijonse.264>
- Anafa, D.M. & Friday, A.C. (2025). 'Design and Evaluation of an Adaptive AI-Powered Learning Companion for Enhancing Digital Literacy in Nigerian Secondary Schools'. *Journal Of Science Innovation and Technology Research*. <https://doi.org/10.70382/ajsitr.v10i9.054>
- Annuš, N., & Kmeť, T. (2024). 'Learn with ME – Let Us Boost Personalized Learning in K-12 Math Education! *Educ. Sci.* 14 (7): 773. <https://doi.org/10.3390/educsci14070773>
- Arnau-González, P., Solera-Monforte, S., Wu, Y., & Arevalillo-Herráez, M. (2025). 'A framework for adapting conversational intelligent tutoring systems to enable collaborative learning'. *Expert Systems with Applications*, 271, 126663. <https://doi.org/10.1016/j.eswa.2025.126663>
- Azevedo, O.B., Morais, F.D., Kautzmann, T.R., & Jaques, P.A. (2020). 'The effects of presenting a worked example before a sequence of isomorphic problems'. In *2020 XV Conferencia Latinoamericana de Tecnologías de Aprendizaje (LACLO)* (pp. 1–8). IEEE. <https://doi.org/10.1109/laclo50806.2020.9381149>
- Bai, Y., Hubers, F., Cucchiarini, C., Hout, R.V., & Strik, H. (2022). 'The Effects of Implicit and Explicit Feedback in an ASR-based Reading Tutor for Dutch First-graders'. *Interspeech 2022*. <https://doi.org/10.21437/interspeech.2022-10810>
- Batsaikhan, B.Z., & Correia, A.-P. (2024). 'The effects of generative artificial intelligence on intelligent tutoring systems in higher education: A systematic review'. *Studies in Technology Enhanced Learning*, 4(1), 1-19. <https://doi.org/10.21428/8c225f6e.33570bb1>
- Bernacki, M.L., Greene, M.J., & Lobczowski, N.G. (2021). 'A systematic review of research on personalized learning: Personalized by whom, to what, how, and

for what purpose(s)?' *Educational Psychology Review*, 33, 1675-1715.
<https://doi.org/10.1007/s10648-021-09615-8>

Bettinger, E., Fairlie, R.W., Kapuza, A., Kardanova, E., Loyalka, P., & Zakharov, A. (2020). 'Diminishing Marginal Returns to Computer-Assisted Learning'.
<http://www.nber.org/papers/w26967>

Bhatt, M.P., Guryan, J., Khan, S.A., Laforest-Tucker, M., Mishra, B., & Ander, R. (2024). 'Can technology facilitate scale? Evidence from a randomized evaluation of high-dosage tutoring'. National Bureau of Economic Research.

Borchers, C., Carvalho, P.F., Koedinger, K.R., & Aleven, V. (2023). 'What Makes Problem-Solving Practice Effective? Comparing Paper and AI Tutoring'. Carnegie Mellon University.

Bringula, R.P., Arlando, F., Atienza, L., Mangubat, J., Roxas, M., Paul, J. et al. (2025). 'Impact of intelligent vs non-intelligent mCSCL on mathematics performance: A comparative study of grade 5 students'. *Acta Psychologica*, 260, 105507. <https://doi.org/10.1016/j.actpsy.2025.105507>

Chinjunthuk, S., Chaiyawut, J., & Junpeng, P. (2025). 'Developing Numbers and Algebra Outcomes Using an Intelligent Tutoring System'. *Journal Of Education and Learning*, 14(6), 233. <https://doi.org/10.5539/jel.v14n6p233>

Clément, B., Sauzéon, H., Roy, D., & Oudeyer, P. (2025). 'Improved Performances and Motivation in Intelligent Tutoring Systems: Combining Machine Learning and Learner Choice'. *arXiv preprint*, arXiv:2402.01669. <https://doi.org/10.48550/arXiv.2402.01669>

Colliot, T., Krichen, O., Girard, N., Anquetil, É., & Jamet, É. (2024). 'What makes tablet-based learning effective? A study of the role of real-time adaptive feedback'. *British Journal of Educational Technology*, 55(5), 2278-2295. <https://doi.org/10.1111/bjet.13439>

Ezzaim, A., Dahbi, A., Haidine, A., & Aqqal, A. (2023). 'Enhancing Academic Outcomes through an Adaptive Learning Framework Utilizing a Novel Machine Learning-Based Performance Prediction Method'. *Data & Metadata*, 2, 164. <https://doi.org/10.56294/dm2023164>

Ezzaim, A., Dahbi, A., Haidine, A., & Aqqal, A. (2024). 'The Impact of Implementing a Moodle Plug-in as an AI-based Adaptive Learning Solution on Learning Effectiveness: Case of Morocco'. *International Journal of Interactive Mobile Technologies*, 18(01), 133-149. <https://doi.org/10.3991/ijim.v18i01.46309>

Feng, M., Brezack, N., Huang, C., Kao, Y., Collins, K., Lee, M. et al. (2025). 'Evaluating the Efficacy of an Intelligent Tutoring System That Integrates Affective Supports Into Math Learning'. *Journal of Computer Assisted Learning*, 41(5). <https://doi.org/10.1111/jcal.70106>

Hallström, J., Mannila, L., Nordlöf, C., Heintz, F., Sperling, K., & Stenliden, L. (2026). 'AI literacy for K–12 education: An international Delphi study'. *Interactive Learning Environments*. Advance online publication. <https://doi.org/10.1080/10494820.2026.2649553>

Haq, I.U., Anwar, A., Basharat, I., & Sultan, K. (2020). 'Intelligent tutoring supported collaborative learning (ITSCCL): A hybrid framework'. *International Journal of Advanced Computer Science and Applications*, 11(8). <https://doi.org/10.14569/IJACSA.2020.0110866>

Hladký, M., & Nagashima, T. (2025). 'The MindfulMath Tutor: A Mindfulness-Infused Learning Environment to Battle Math Anxiety'. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. ACM <https://doi.org/10.1145/3706599.3719922>.

Holmes, W., Bialik, M., & Fadel, C. (2019). 'Artificial intelligence in education: Promises and implications for teaching and learning'. Boston, MA: Center for Curriculum Redesign. <https://curriculumredesign.org/our-work/artificial-intelligence-in-education/>

Horvers, A., Kooi, R., Campen, C.A.N.K., Dijkstra, R., Baars, M. & Molenaar, I. (2024). 'How does co-regulation with Adaptive Learning Technologies affect primary school students' goal-setting, regulation of practice behavior and learning outcomes?' *Frontiers in Education*, 9. <https://doi.org/10.3389/feduc.2024.1435483>

Huang, Y. (2025). 'Optimizing Student Engagement: Reinforcement Learning Approaches to Problem Selection in ITS'. *Applied And Computational Engineering*, 210(1), 105-119. <https://doi.org/10.54254/2755-2721/2025.Id30332>

Huang, Y., Aleven, V., McLaughlin, E. & Koedinger, K. (2020). 'A General Multi-method Approach to Design-Loop Adaptivity in Intelligent Tutoring Systems'. *Lecture Notes in Computer Science*. In *Artificial Intelligence in Education: AIED 2020*, Lecture Notes in Computer Science, 12164, 124–129. Springer. https://doi.org/10.1007/978-3-030-52240-7_23

Huang, Y., Lobczowski, N.G., Richey, J.E., Mclaughlin, E.A., Asher, M.W., Harackiewicz, J.M. et al. (2021). 'A General Multi-method Approach to Data-Driven Redesign of Tutoring Systems'. In *Proceedings of the Eighth ACM Conference on Learning @ Scale (L@S '21)*. ACM. <https://doi.org/10.1145/3448139.3448155>

Jaloliddin, J.S. (2025). 'AI-Augmented Gamified Learning And Its Impact On Motivation In Primary English Classes'. *International Journal of Pedagogics*, 5(10), 58-61. <https://doi.org/10.37547/ijp/volume05issue10-11>.

Katchapakirin, K., Anutariya, C., & Supnithi, T. (2022). 'ScratchThAI: A conversation-based learning support framework for computational thinking

development'. *Education And Information Technologies*, 27(6), 8533-8560. <https://doi.org/10.1007/s10639-021-10870-z>

Khazanchi, R., Mitri, D.D., & Drachsler, H. (2024). 'The Effect of AI-Based Systems on Mathematics Achievement in Rural Context: A Quantitative Study'. *Journal Of Computer Assisted Learning*, 41(1). <https://doi.org/10.1111/jcal.13098>

Liu, Z., Lin, G., Li, H., Zhang, H., Lu, Y., Gao, X. et al. (2025). 'SingaKids: A Multilingual Multimodal Dialogic Tutor for Language Learning'. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)* (pp. 1244–1253). Association for Computational Linguistics. <https://arxiv.org/abs/2506.02412v1>

Ma, W., Adesope, O.O., Nesbit, J. C., & Liu, Q. (2014). 'Intelligent tutoring systems and learning outcomes: A meta-analysis'. *Journal of Educational Psychology*, 106(4), 901-918. <https://doi.org/10.1037/a0037123>

Maaz, N., Mounsef, J. & Maalouf, N. (2025). 'CARE: Towards customized assistive robot-based education'. *Frontiers in Robotics and AI*, 12, Article 1474741. <https://doi.org/10.3389/frobt.2025.1474741>

Maity, S., & Deroy, A. (2024). 'Generative AI and its impact on personalized intelligent tutoring systems'. *arXiv*. <https://doi.org/10.48550/arXiv.2410.10650>

Majumdar, R., Takami, K., & Ogata, H. (2023). 'Learning with Explainable AI-Recommendations at School: Extracting Patterns of Self-directed Learning from Learning Logs'. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 245–249). IEEE. <https://doi.org/10.1109/icalt58122.2023.00078>

Marwiang, M., Prasertsang, M., & Junpeng, P. (2025). 'Enhancing Students' Learning Outcomes in Mathematics through Intelligent Tutoring Systems Based on Real-Time Feedback'. *Journal Of Education and Learning*, 14(6), 186. <https://doi.org/10.5539/jel.v14n6p186>

Matayoshi, J., Uzun, H., & Cosyn, E. (2020). 'Studying Retrieval Practice in an Intelligent Tutoring System'. In *Proceedings of the Seventh ACM Conference on Learning @ Scale (L@S '20)*. ACM. <https://doi.org/10.1145/3386527.3405927>

Maurya, K.K., Srivatsa, K.V.A., Petukhova, K., & Kochmar, E. (2025). 'Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors'. In L. Chiruzzo, A. Ritter, & L. Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Vol. 1, pp. 1234-1251). <https://doi.org/10.18653/v1/2025.naacl-long.57>

- Meeter, M. (2021). 'Primary school mathematics during the COVID-19 pandemic: No evidence of learning gaps in adaptive practicing results'. *Trends in Neuroscience and Education*, 25, 100163. <https://doi.org/10.1016/j.tine.2021.100163>
- Merrill, M.D. (2002). 'First principles of instruction'. *Educational Technology Research and Development*, 50(3), 43-59. <https://doi.org/10.1007/BF02505024>
- Mondia, D.J., Naranjo, K.N., Galamay, S.T., & Tiglao, N.M.C. (2023). 'AIComprehend: An Adaptive Reading Comprehension Learning Platform using Machine Learning'. In *2023 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE. <https://doi.org/10.1109/isncc58260.2023.10323847>
- Morais, F.D., & Jaques, P.A. (2022). 'Does Handwriting Impact Learning on Math Tutoring Systems?' *Informatics in Education*, 1: 55-90. <https://doi.org/10.15388/infedu.2022.03>
- Munshi, A., Biswas, G., Baker, R., Ocumpaugh, J., Hutt, S., & Paquette, L. (2022). 'Analysing adaptive scaffolds that help students develop self-regulated learning behaviours'. *J. Comput. Assist. Learn*, 39: 351-368). <https://doi.org/10.1111/jcal.12761>
- Ningsih, A.G., Atmazaki & Ramadhan, S. (2026). 'AI-Powered Adaptive Quizzing: Enhancing Personalized Learning, Student Engagement, and Performance in Digital Classrooms. Salud Ciencia Y Tecnología'. *Salud Ciencia Y Tecnología*. 6, 2647. <https://doi.org/10.56294/saludcyt20262647>
- Nwana, H.S. (1990). 'Intelligent tutoring systems: An overview'. *Artificial Intelligence Review*, 4(4), 251-277. <https://doi.org/10.1007/BF00168958>
- Obermüller, F., Greifenstein, L., & Fraser, G. (2023). 'Effects of Automated Feedback in Scratch Programming Tutorials.' In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education*, Vol. 1 (pp. 396–402). ACM. <https://doi.org/10.1145/3587102.3588803>
- OECD (2025). *Results from TALIS 2024: The state of teaching*. OECD Publishing. <https://doi.org/10.1787/90df6235-en>
- OECD (2026). *OECD Digital Education Outlook 2026: Exploring effective uses of generative AI in education*. OECD Publishing. <https://doi.org/10.1787/062a7394-en>
- Olmo-Muñoz, J.D., José, A., González-Calero, Diago, P.D., Arnau, D., & Arevalillo-Herráez, M. (2022). 'Using intra-task flexibility on an intelligent tutoring system to promote arithmetic problem-solving proficiency'. *British Journal of Educational Technology*, 53(6), 1976-1992. <https://doi.org/10.1111/bjet.13228>

Olney, A.M., D'Mello, S.K., Person, N., Cade, W., Hays, P., Dempsey, C.W., Lehman, B., Williams, C., & Graesser, A.C. (2025). 'Efficacy of a computer tutor that models expert human tutors'. In *Proceedings of the 26th International Conference on Artificial Intelligence in Education (AIED 2025)*. Springer. https://doi.org/10.1007/978-3-031-98462-4_38

Oreopoulos, P., Gibbs, C., Jensen, M., & Price, J. (2024). '*Teaching Teachers to Use Computer Assisted Learning Effectively: Experimental and Quasi-Experimental Evidence*'. NBER Working Paper No. 32388. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w32388/w32388.pdf

Oreopoulos, P., Keyes-Krysakowski, O., & Agarwal, D. (2026). 'How In-School Supervised Ed-Tech Support Produces Massive Learning Gains: A Khan Academy Field Experiment in India'. National Bureau of Economic Research. <http://www.nber.org/papers/w34683>

Pervaiz, Z., Bakar, A., & Saddique, S. (2025). 'Enhancing Learning Outcomes through AI-Based Tutoring Systems: A Study on Student Motivation and Academic Achievement'. *ACADEMIA International Journal for Social Sciences*, 4(3), 5427–5449. <https://doi.org/10.63056/acad.004.03.0805>

Prasertsang, M., Marwiang, M., & Junpeng, P. (2025b). 'Enhancing mathematical conceptual structure in number and algebra through intelligent tutoring systems for Demonstration School of Khon Kaen University Thailand'. In *2025 40th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. IEEE. <https://doi.org/10.1109/itc-csc66376.2025.11137723>

Prasertsang, M., Marwiang, M., & Junpeng, P. (2025a). 'Developing Mathematical Procedures in Numbers and Algebra Through an Intelligent Tutoring System'. *Journal Of Education and Learning*, 14(6), 245. <https://doi.org/10.5539/jel.v14n6p245>

Sayed, W.S., Noeman, A.M., Abdellatif, A., Abdelrazek, M., Badawy, M.G., Hamed, A. et al. (2022). 'AI-based adaptive personalized content presentation and exercises navigation for an effective and engaging E-learning platform'. *Multimedia Tools and Applications*, 82(3), 3303-3333. <https://doi.org/10.1007/s11042-022-13076-8>

Shih, S.C., Tsai, H.Y., & Chen, M.L. (2021). 'The Effect of a One-On-One Dialogue-Based Mathematical Intelligent Tutoring System for Learning Equivalent Fraction'. In *Education and New Developments 2021* (pp. 363–365). <https://doi.org/10.36315/2021end077>

Shih, S., Chang, C., Kuo, B., & Huang, Y. (2023). 'Mathematics intelligent tutoring system for learning multiplication and division of fractions based on diagnostic

teaching'. *Education And Information Technologies*, 28(7), 9189-9210. <https://doi.org/10.1007/s10639-022-11553-z>

Simbolon, M., Pongkondek, J.J., Henukh, A., & Ronchintaniawati, D. (2025). 'AI-Driven Sociocultural Interactive Digital Module for Papua: Advancing Educational Technology to Sustainable Development Goals'. *International Journal of Learning Teaching and Educational Research*, 24(2), 543-559. <https://doi.org/10.26803/ijlter.24.2.27>

Sottolare, R., Graesser, A., Hu, X., & Goldberg, B. (2014). *Design recommendations for intelligent tutoring systems: Volume 2. Instructional management*. U.S. Army Research Laboratory.

Standen, P.J., Brown, D.J., Taheri, M., Trigo, M.J.G., Boulton, H., Burton, A. et al. (2020). 'An evaluation of an adaptive learning system based on multimodal affect recognition for learners with intellectual disabilities'. *British Journal of Educational Technology*, 51(5), 1748-1765. <https://doi.org/10.1111/bjet.13010>

Su, M., Dang, B., Nguyen, A., & Nagashima, T. (2025). 'Choice-making in an adaptive learning system with motivational pedagogical agents'. *NPJ Science of Learning*. 10, Article 77. <https://doi.org/10.1038/s41539-025-00366-7>

Thampitak, D., & Junpeng, P. (2025). 'Developing the Structure of Observed Learning Outcomes in the Measurement and Geometry Strand for Grade 8th Gifted Students through Intelligence Tutoring Systems'. In *2025 40th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*. IEEE. <https://doi.org/10.1109/itc-csc66376.2025.11137645>

Thomas, D.R., Gatz, E., Gupta, S., Aleven, V., & Koedinger, K.R. (2024a). 'The Neglected 15%: Positive Effects of Hybrid Human-AI Tutoring Among Students with Disabilities'. Carnegie Mellon University.

Thomas, D.R., Lin, J., Gatz, E., Gurung, A., Gupta, S., Norberg, K. et al. (2024b). 'Improving Student Learning with Hybrid Human-AI Tutoring: A Three-Study Quasi-Experimental Investigation'. In *Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24)*. ACM. <https://doi.org/10.1145/3636555.3636896>

Uriarte-Portillo, A., Zatarain-Cabada, R., Barrón-Estrada, M.L., Ibáñez, M.B., & González-Barrón, L.-M. (2023). 'Intelligent Augmented Reality for Learning Geometry'. *Information*, 14, 245. <https://doi.org/10.3390/info14040245>

VanLehn, K. (2011). 'The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems'. *Educational Psychologist*, 46(4), 197-221. <https://doi.org/10.1080/00461520.2011.611369>

Verma, A., & Kulkarni, R. (2025). 'Adaptive Learning Algorithms for Differentiated Instruction in Mathematics Education'. *International Academic Journal of Science and Engineering*, 12(3), 50-53. <https://doi.org/10.71086/iajse/v12i1/iajse1209>

Vogel, C. (2023). *Expertise – Bots – Fernlehre: Eine gelungene Kombination? Eine Design-Based-Research-Studie zur Förderung von Adaptive Expertise durch den Einsatz eines Educational Bots in der Fernlehre* [doctoral dissertation]. FernUniversität in Hagen. https://ub-deposit.fernuni-hagen.de/servlets/MCRFileNodeServlet/mir_derivate_00002563/Diss_Vogel_Expertise_Bots_Fernlehre_2023.pdf

Wang, F., Zhou, X., Li, K., Cheung, A.C.K., & Tian, M. (2025). 'The effects of artificial intelligence-based interactive scaffolding on secondary students' speaking performance, goal setting, self-evaluation, and motivation in informal digital learning of English'. *Interactive Learning Environments*, 33(7), 4633-4652. <https://doi.org/10.1080/10494820.2025.2470319>

Wang, X., Neitzel, A.J., & Madden, N. A. (2024). 'Lightning Squad: Assessing the Dosage Effect of Computer-Assisted Tutoring with Cooperative Learning for Struggling Readers'. *Journal of Education for Students Placed At Risk (JESPAR)*, 1–19. <https://doi.org/10.1080/10824669.2024.2388242>

Wen, X., & Pan, D. (2025). 'Charting new territories: fuzzy systems in English language teaching and learning'. *Peer J. Computer Science*, 11, e2887. <https://doi.org/10.7717/peerj-cs.2887>

Wijekumar, K., Lei, P., Rice, M., Beerwinkle, A., Zhang, S., & Meyer, B.J.F. (2024). 'A Web-Based Intelligent Tutoring System for Reading Comprehension Delivered to Fifth-Grade Students Attending High-Poverty Schools: Results From a Replication Efficacy Study'. *Journal of Educational Psychology*, 116(8), 1333-1351. <https://doi.org/10.1037/edu0000878>

Wijekumar, K., Meyer, B.J., Puiwa, Lei, Beerwinkle, A.L., & Joshi, M. (2019). 'Supplementing teacher knowledge using web-based Intelligent Tutoring System for the Text Structure Strategy to improve content area reading comprehension with fourth- and fifth-grade struggling readers'. *Dyslexia*, 26(2), 120-136. <https://doi.org/10.1002/dys.1634>

Wutthisarn, W., & Junpeng, P. (2024). 'The Effect of a Diagnostic System for Scientific Competency Combined with a Personalized Intelligent Tutoring Module through Machine Learning for Enhancing the Learning Progression of Seventh Grade Students'. *International Journal of Religion*, 5(11), 6192-6208. <https://doi.org/10.61707/mwsxkt90>

Yan, J. (2024). 'Effectiveness Analysis and Optimization Path of AI Tutoring Models in Foreign Language Learning Applications'. *Applied Mathematics and Nonlinear Sciences*, 9(1). <https://doi.org/10.2478/amns-2024-1341>.

Yannier, N., Hudson, S.E., Chang, H., & Koedinger, K.R. (2024). 'AI Adaptivity in a Mixed-Reality System Improves Learning'. *International Journal of Artificial Intelligence in Education*, 34(4), 1541-1558. <https://doi.org/10.1007/s40593-023-00388-5>

Yu, H., Alessio, D.A., Lee, W., Rebelsky, W., Sylvia, F., Murray, T. et al. (2023). 'COVES: A Cognitive-Affective Deep Model that Personalizes Math Problem Difficulty in Real Time and Improves Student Engagement with an Online Tutor'. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. ACM. <https://doi.org/10.1145/3581783.3613965>

Zhai, J., Sudiarta, G.P., Santosa, M.H., Wayan, I., & Astawa, P. (2025). 'Research on intelligent regulation mechanisms of learner cognitive load in digital learning environments'. *Future Technology*, 4(4), 205-215. <https://doi.org/10.55670/fpll.futech.4.4.17>

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct centres. You can find the address of the centre nearest you online (european-union.europa.eu/contact-eu/meet-us_en).

On the phone or in writing

Europe Direct is a service that answers your questions about the European Union. You can contact this service:

- by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
- at the following standard number: +32 22999696,
- via the following form: european-union.europa.eu/contact-eu/write-us_en.

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website (european-union.europa.eu).

EU publications

You can view or order EU publications at op.europa.eu/en/publications. Multiple copies of free publications can be obtained by contacting Europe Direct or your local documentation centre (european-union.europa.eu/contact-eu/meet-us_en).

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex (eur-lex.europa.eu).

EU open data

The portal data.europa.eu provides access to open datasets from the EU institutions, bodies and agencies. These can be downloaded and reused for free, for both commercial and non-commercial purposes. The portal also provides access to a wealth of datasets from European countries.



Publications Office
of the European Union