frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 21 January 2026
DOI 10.3389/feduc.2025.1738655

Check for updates

OPEN ACCESS

EDITED BY
Huma Akram,
North China University of Water Resources and Electric Power, China

REVIEWED BY
Abbas Hussein Abdelrady,
Qassim University, Saudi Arabia
Iden Rainal Ihsan,
Universitas Samudra, Indonesia

*CORRESPONDENCE
Julius Schaaf
✉ schaaf@math.uni-frankfurt.de

RECEIVED 03 November 2025
REVISED 23 December 2025
ACCEPTED 29 December 2025
PUBLISHED 21 January 2026

CITATION
Schaaf J, Rolfes T, Nagy G and Heinze A (2026) The effect of the frequency of use of an intelligent tutoring system on learning gains in mathematics secondary education.
Front. Educ. 10:1738655.
doi: 10.3389/feduc.2025.1738655

COPYRIGHT
© 2026 Schaaf, Rolfes, Nagy and Heinze. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

The effect of the frequency of use of an intelligent tutoring system on learning gains in mathematics secondary education

Julius Schaaf[1]*, Tobias Rolfes[1], Gabriel Nagy[2] and Aiso Heinze[3]

[1]Institute of Mathematics and Computer Science Education, Faculty of Computer Science and Mathematics, Goethe University, Frankfurt am Main, Germany, [2]Educational Measurement and Data Science, Leibniz Institute for Science and Mathematics Education, Kiel, Germany, [3]Mathematics Education, Leibniz Institute for Science and Mathematics Education, Kiel, Germany

Intelligent tutoring systems (ITS) are characterized by their direct and adaptive feedback as well as their capability of assessing the knowledge of students and administering exercises based on those assessments. The state of research regarding the effectiveness of ITS in mathematics is inconclusive. Hence, in this study, we examined the impact of utilizing an ITS on the learning gains in mathematics for students in grades 7 and 8. This longitudinal investigation was conducted with students from 55 classes (940 students) in northern Germany. Mathematics performance as well as relevant covariates were measured at the beginning and the end of the school year, and interactions with the ITS were recorded throughout the school year, providing a comprehensive dataset for analysis. Teachers were free in choosing the extent, subjects and methods of ITS usage. In addition, students could also use the ITS on their own. A multilevel analysis revealed that the frequency of ITS usage had no significant effect, neither at the class level nor at the individual level. Our results show that using ITS does not automatically lead to better learning gains. Therefore, future studies need to identify the conditions and practices that contribute to effective ITS use.

KEYWORDS

empirical study, Germany, ITS, longitudinal study, mathematics education

1 Introduction

Education is becoming increasingly digital. Although digital media have been used in schools for several decades, school closings during the COVID-19 pandemic have inevitably accelerated this trend considerably. At the same time, education is expected to become more individualized. Digital media could be seen as a tool to implement individualized opportunities to learn, and to improve students' learning processes. This is especially true for *intelligent tutoring systems* (ITS). They are digital systems designed to enable effective learning through feedback and individualized task delivery based on students' prior performance (VanLehn, 2006).

Evidence from several meta-analyses indicates that these systems improve student outcomes across a range of contexts (e.g., Kulik and Fletcher, 2016; Steenbergen-Hu and Cooper, 2014; Ma et al., 2014; Higgins et al., 2012). One of the key benefits of ITS lies in their ability to provide adaptive feedback and facilitate self-paced learning, both of which are grounded in well-established and empirically supported educational theories (Hillmayr et al., 2020).

While ITS generally lead to positive learning gains, the effectiveness of these systems, particularly in mathematics, is more varied and less definitive. Research by Kulik and Fletcher

(2016) shows that the impact of ITS in mathematics is smaller than in other subjects, and a study by Steenbergen-Hu and Cooper (2013) indicates a minimal effect when using ITS for mathematics. Ma et al. (2014), in contrast, find no significant differences in the effectiveness of ITS between subjects. Despite this mixed evidence, commercial ITS platforms are becoming increasingly widespread in educational settings. In Germany, commercial ITS such as MatheGym and Bettermarks report hundreds of thousands of users (MatheGym, 2025; Bettermarks, 2025). Considering these trends, we aimed to evaluate whether the use of ITS correlates with improved learning outcomes in mathematics. To assess this, we conducted a longitudinal study with both pre- and posttests, measuring students' mathematics performance with a curriculum based standardized test alongside key factors such as their attitudes toward mathematics and levels of mathematics anxiety. Additionally, aggregated log data from the ITS was collected. We analyzed the data using a Rasch model and applied multilevel statistical modeling to gain deeper insights into the relationship between ITS usage and learning gains.

2 Theoretical framework

In this section, we provide a theoretical rationale for the potential benefits of digital tools in promoting more effective learning. Starting with key educational principles and a definition of ITS, we review the current state of research on the effectiveness of ITS in education. Finally, we summarize existing research on the use of ITS specifically for mathematics learning.

2.1 Digital tools

According to Hattie (2023) meta-analysis, the use of digital tools has, on average, a moderate effect on learning gains ($d = 0.34$). Beyond cognitive outcomes, digital learning environments have also been shown to influence affective dimensions of learning. For instance, prior research suggests that such environments may help reduce students' test anxiety (Akram and Abdelrady, 2023). In addition, interactive digital tools can enhance students' self-efficacy, perceived instructional clarity, and learning expectations (Akram and Abdelrady, 2025), indicating that their benefits extend to motivational and experiential aspects of learning. Taken together, these findings suggest that digital media can support students' learning in multiple ways. However, the effectiveness of digital tools depends strongly on how they are implemented. Simply introducing technology into the classroom does not automatically lead to improved performance (Hattie, 2023). To ensure that digital media contribute meaningfully to learning, their use should be aligned with four well-established educational principles: multimedia learning, self-paced learning, guided activity, and feedback (Hillmayr et al., 2020). Multimedia learning is conceptualized in the *cognitive theory of multimedia learning* (CTML; Mayer, 2014). This theory asserts three underlying assumptions. First, humans process information via two separate channels: the visual pictorial and the auditive verbal channel. Second, both channels only have limited, but separate bandwidths, meaning the speed at which information can be processed is limited. Third, currently presented content must be engaged with and processed actively. Active engagement with content is defined via active learning "which entails carrying out a coordinated set of cognitive processes during learning (i.e., active processing assumption)" (Mayer, 2014, p. 43). The first two assumptions explain the benefit of using multimedia in education, since more information can be processed if both channels are used without causing cognitive overload. The third assumption suggests that many digital tools can facilitate learning because they encourage students to actively engage with the content through interactive learning environments.

Furthermore, self-paced learning is considered to be an important part of effective learning because it enables students to progress through topics at their own pace (Hillmayr et al., 2020; Moreno, 2007). Students perceive tasks as less difficult if they can control the speed at which they work on them (Moreno, 2007). It has been shown that students perform better when self-pacing than following prespecified pacing, even when controlling for the total study-time (Tullis and Benjamin, 2011).

Additionally, guided activity holds significant value in interactive digital educational settings. Students find tasks enriched with guided activities less demanding than those without (Moreno and Mayer, 2007). Belland et al. (2017) concluded in their meta-analysis that scaffolding, a form of guided activity, has a significant positive effect (g[1] = 0.46) on learning outcomes in STEM education. Encouraging students to actively engage in selecting, organizing, and integrating new information fosters essential and generative processing, thereby enhancing learning outcomes (Moreno and Mayer, 2007).

Finally, feedback is an important part of effective learning processes and has been proven to enhance learning gains (Hattie and Timperley, 2007). Feedback can be categorized in different ways. Two common categorizations include differentiation by timing (*immediate* vs. *delayed* feedback), and comprehensiveness (*correct response* vs. *elaborated* feedback). Immediate feedback has been shown to be more beneficial than delayed feedback in applied settings such as classrooms ($d = 0.28$; Kulik and Kulik, 1988). Van Van Der Kleij et al. (2015) showed that elaborated feedback results in larger effect sizes ($g = 0.49$) than feedback only consisting of correct responses ($g = 0.32$) or feedback only consisting of information whether the answer given was correct ($g = 0.05$).

As interactive learning environments enable students to engage in multimedia, individualized, self-paced learning processes with direct feedback, they are expected to facilitate effective learning and enhance learning outcomes.

2.2 Intelligent tutoring systems

An example of a digital tool designed to incorporate all these features —multimedia, self-pacing, guided activity, elaborated feedback—are ITS. The following section provides an overview of ITS and the current state of research regarding their effectiveness.

Computer tutoring systems in schools have been around for almost 60 years (Atkinson, 1968). They are generally divided into two generations. The first generation is known as *computer assisted instruction* (CAI) (Kulik and Fletcher, 2016; VanLehn, 2011), while

1 Hedges' g is a standardized effect size used for comparing groups of unequal sizes and can be interpreted similarly to Cohen's d.

systems from the second generation are usually called ITS (VanLehn, 2011). The main difference between CAI and ITS lies in the level of adaptivity and the types of scaffolding provided. While there are various definitions of ITS (e.g., Ma et al., 2014; Shute and Zapata-Rivera, 2007), most emphasize that, unlike CAI, ITS not only offer corrective feedback and hints but also allow students to choose their approach. Additionally, ITS provide adaptive feedback and feedback for intermediate steps, known as sub-step feedback (VanLehn, 2006, 2011). VanLehn calls this behavior the *inner loop*. The *outer loop*, in contrast, governs the type and difficulty of tasks a student is presented with within the system. In practice, this means ITS can, for example, assign more challenging tasks to well-performing students while directing underperforming students to review fundamental topics.

Ma et al. (2014) have summarized the three features an ITS needs in order to conduct these loops properly:

"An ITS is a computer system that for each student:

1 Performs tutoring functions by (a) presenting information to be learned, (b) asking questions or assigning learning tasks, (c) providing feedback or hints, (d) answering questions posed by students, or (e) offering prompts to provoke cognitive, motivational, or metacognitive change.
2 By computing inferences from student responses constructs either a persistent multidimensional model of the student's psychological states (such as subject matter knowledge, learning strategies, motivations, or emotions) or locates the student's current psychological state in a multidimensional domain model.
3 Uses the student modeling functions identified in point 2 to adapt one or more of the tutoring functions identified in point 1." (p. 902).

2.2.1 Effectiveness of ITS

The effect size reported for ITS in education is substantially influenced by both study design and subject matter. Specifically in mathematics, the body of research remains inconclusive.

A meta-analysis conducted by VanLehn (2011) found an effect size of $d = 0.76$, which is substantially larger than the effect sizes for CAI (*Glass' ES*[2] = 0.35; Tamim et al., 2011) or the general usage of technology, ($d = 0.34$; Hattie, 2023, 293). Other meta-analyses find similar values. Kulik and Fletcher (2016) report a median Glass' *ES* of 0.66, Steenbergen-Hu and Cooper (2014) report values of Hedges' g between $g = 0.35$ and 0.37. Ma et al. (2014) report a mean effect size of $g = 0.41$.

The effectiveness of ITS varies across studies, influenced by multiple factors such as teacher proficiency, control group composition, instructional methods, assessment types, study duration, and subject domain. These factors will be discussed in more detail in the following.

The teacher's proficiency with the ITS plays a significant role. Koedinger and Anderson (1993) reported substantial differences in the effectiveness of the ITS called "Angle" based on how much experience the teachers had in using this specific ITS. They reported an effect size of Glass' $ES = 0.96$ for experts and −0.23 for novices. This indicates that the unfavorable use of digital tools by teachers can even impede students' understanding of the topic. One possible explanation is how the ITS is used. The expert employed Angle as a supplementary resource, whereas the novice relied on it as the primary instructional method. Leaving students working with the ITS for extended periods of time during lessons could result in reduced student-teacher interactions, leading students to feel isolated and consequently less motivated. Furthermore, even if ITS are used supplementary (e.g., for homework) they could have adverse effects if not properly implemented. Hattie (2023) highlights the importance of integrating homework into the school's curriculum for it to be effective. If teachers stop grading or discussing homework because the ITS provides feedback, students may perceive their work as undervalued.

Next, to an adequate implementation, the effect size measured depends on the nature of the control group (Ma et al., 2014). Specifically, when comparing the use of ITS to non-ITS computer-based instruction, the average effect size was $g = 0.57$. In contrast, when ITS use was compared to small group instruction without digital tools, the average effect size was $g = -0.11$. So, while ITS seem to be more effective than CAI, they seem to be less effective than human tutoring in small groups. In addition, ITS was more effective as a supplementary tool than as the primary method of instruction (see Ma et al., 2014, p. 909).

Additionally, Kulik and Fletcher (2016) report a difference based on the type of test used to measure student achievement: Evaluations of the cognitive tutor, a prominent American ITS, consistently showed a positive significant effect (*Glass' ES* = 0.73) when evaluated with tests that were developed to measure the type of tasks focused on in the program. However, when using standardized tests, no significant effect was found ($ES = 0.13$[3]). Evaluations using a combination of both types of tests found an average effect of $ES = 0.45$.

Furthermore, shorter interventions also seem to increase the effect size (Kulik and Fletcher, 2016). Therefore, studies with long measurement periods have small expected effect sizes. Nevertheless, studies measuring the effect over an entire school year provide a realistic estimate of the learning gains expected from long-term use.

The subject domain also plays a crucial role. Although the positive effect of ITS is empirically well-supported in several domains, the situation is less clear for ITS in mathematics. Kulik and Fletcher (2016) demonstrated that the use of ITS in mathematics has a significantly smaller effect size than in other subjects. However, in this meta-analysis, studies focused on mathematics generally also had larger sample sizes and employed standardized tests, which are typically associated with smaller effect sizes. Steenbergen-Hu and Cooper (2013) conducted a meta-analysis of ITS used exclusively in mathematics and found no significant learning gains for the use of ITS in mathematics ($g = 0.01$ to $g = 0.09$). Ma et al. (2014) observed that ITS used in humanities and social sciences had significantly higher effect sizes than those used in mathematics and science, while Higgins et al. (2012), on average, reported a greater effect for ITS in mathematics and science than other fields such as literacy. In summary, the body of

2 *Glass' ES* is a standardized effect size used to compare groups with unequal variances and can be interpreted similarly to Cohen's *d*.

3 Although the effect was not statistically significant, we report this value for comparison with other findings.

research on the effectiveness of ITS for mathematics learning is inconclusive.

3 Current study

Despite the uncertainties regarding empirical effectiveness mentioned in the previous section, the use of ITS in mathematics instruction is widespread. In the US, cognitive tutor, a popular ITS, and its successor, MATHia are used by hundreds of thousands of students (U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse, 2009), even though, learning with the ITS does not seem to increase performance in standardized tests, as mentioned earlier. In Germany, MatheGym, and Bettermarks are two very prominent ITS. MatheGym has approximately 150,000 users (as of April, 2024) (MatheGym, 2025). Bettermarks is used by 500,000 students according to self-reported numbers by the company (Bettermarks, 2025). In addition, at least 7 out of 16 federal states in Germany have state-wide licenses for Bettermarks. Therefore, it is a relevant question how effective these systems are in practice. Due to its broad use in Germany, we selected the Bettermarks system to investigate the effectiveness of ITS in supporting mathematics learning.

3.1 Research question

We opted to analyze the effect in a real, ecologically valid setting, where individual teachers had the freedom to determine the extent of ITS usage in their respective classes. This design was chosen because it allowed to investigate the effect sizes of ITS usage in real-world conditions. As mentioned above, the effectiveness of ITS depends on many different factors—some of which may differ substantially between experimental and field settings—including factors beyond our control. Therefore, we chose to observe the implementation of an ITS in an authentic educational context, and, more specifically, to analyze how the frequency of ITS use is related to learning outcomes.

If the ITS is effective, we would expect students who use it more frequently to show greater learning gains in mathematics performance, even when controlling for covariates relevant to learning. Since the system is designed to support students in mastering curricular content, these learning gains will be assessed based on students' performance on standard curricular tasks. Hence, our research question is:

> *RQ*: What is the relationship between the frequency of ITS use and student learning gains in curricular mathematics tasks?

As mentioned above, on the one hand, there are reasons to expect a positive effect of the ITS due to the effectiveness of multimedia learning, adaptive and immediate feedback, adaptive tasks, and self-paced learning. On the other hand, reduced direct teacher-student interactions and the resulting reduced social support lead to an expectation of a negative effect, as does the constraint to solve tasks in a specific way dictated by the ITS through its scaffolding. Therefore, it is not clear whether the use of ITS in learning mathematics leads to a higher or lower performance gain.

4 Methods

4.1 Design

The study was conducted as a one-year longitudinal study with pre- and posttest. In order to investigate authentic ITS use, the teachers were not given any regulations regarding ITS use. They were free to use the system in their classes as much or as little as they deemed appropriate. Consequently, substantial variance in ITS usage was anticipated. In our design, this natural variability in ITS usage was used to estimate the expected effect of ITS usage on learning gains under common conditions.

At the beginning and end of the school year, a computer-based questionnaire and test was administered. The pretest was conducted between September and November 2021, while the posttest took place in June or July 2022. The exact timing of test administration was determined by the respective teachers, resulting in varying time intervals between pre- and posttests across classes. The average time interval between pre- and posttest was $M = 250$ days ($SD = 21$). Mathematics performance was measured in both tests. Given the correlational nature of the study design, a comprehensive set of student-related variables was recorded in order to control for individual differences and thus enable the estimation of the effect of ITS usage (see Section 3.4 Instruments).

Furthermore, throughout the school year, the activity of individual students in the ITS was automatically recorded in the form of aggregated data. Students completed so-called *worksheets* within the ITS. Each worksheet comprised a set of similar tasks on a common topic. Each time a worksheet was completed by a student, one datapoint was collected, containing three pieces of information: The number of tasks on the worksheet, the number of tasks solved correctly, and the start time of the work on the worksheet. For technical reasons, more detailed data could not be collected.

4.2 Sample

The study was conducted accompanying the rollout of a test license for the used ITS in three counties in the German federal state Schleswig-Holstein. All schools—and by extension, all students—in those counties had free access to the ITS, regardless of their participation in the study. The study was conducted with 7th and 8th grade students from 82 classes across 13 schools. The pretest was completed by 1,673 students, while the posttest was completed by 1,309 students. In total, 1,062 students were successfully matched with both a pretest and a posttest. Of these students, 120 could not be linked to an account in the ITS and were therefore excluded from the longitudinal analysis. Additionally, two more students had to be removed because they were the only participants in their respective classes with parental consent, which would have caused complications during the analysis. Thus, data from 940 students across 57 classes were available for longitudinal analysis. However, data from students with only one available measurement point were still used when feasible (e.g., scaling the Rasch model).

The average age of the sample used for longitudinal analysis at the pretest was 12.8 years, with a standard deviation of 0.7 years. Of these, 489 were female, 445 were male, and 8 did not specify their gender.

4.3 ITS Bettermarks

The ITS Bettermarks is a digital learning platform designed to align with the state-level mathematics curricula of German secondary education and can be used both in the classroom and as a supplementary tool (Bettermarks, 2025). The ITS is intended to support the assessment and development of students' competencies within the curricular frameworks governing secondary mathematics education in Germany (Bettermarks, 2025). Therefore, it offers a comprehensive range of tasks covering the prescribed content of secondary education in Germany. The material provided by the ITS for Grades 7 and 8 shows substantial correspondence with the TIMSS framework (Mullis et al., 2021), aligning largely with the content domains Number, Algebra, and Geometry and Measurement, as well as with the cognitive domains Knowing and Reasoning.

Access to the platform is available to both students and teachers via a browser or an app. The content within the ITS is organized in a folder structure (Figure 1). Each topic has its own *book*, comprising an introduction, several *chapters*, a test to assess understanding of the covered content, and a review section for students to revisit the topic and assess their proficiency. Each chapter consists of various *worksheets*, each containing multiple *tasks*.

Teachers can assign digital worksheets, known as *to do's*, to their students, allowing them to give digital homework, for example. Teachers can view the completion rates of individual students as well as their inputs. Students can also independently select tasks from all books and work on them autonomously.

Most tasks in the worksheets focus on promoting procedural knowledge, but there are also tasks that promote conceptual knowledge (e.g., Hiebert and Lefevre, 1986). Tasks often involve performing common standardized procedures, such as adding fractions, determining a function equation, or drawing and calculating the area of a triangle. The ITS contains few tasks that encourage constructive reasoning. As a result, input formats often consist of numbers or fractions. Sometimes more interactive inputs are required, such as digitally drawing a function graph using a toolbox or iconically representing a fraction by dividing and coloring an area.

More complex tasks are broken down into substeps. For example, the breakdown of adding two unlike fractions is as follows:

a Bring both fractions to the lowest common denominator.
b Add the two fractions.
c Simplify and convert the result into a mixed number.

These substeps are worked on sequentially. The ITS provides direct feedback for each input. Some of the feedback is corrective ("Sorry, that's not right."), while other feedback is adaptive ("You have found a common denominator, but the lowest common denominator is smaller."). If an incorrect input is made twice, this substep is marked as incorrect, and the solution to this substep is shown as a worked example. These features correspond to the inner loop according to VanLehn (2006, 2011). Additionally, the ITS recognizes when students consistently make specific types of errors. For instance, if a student demonstrates proficiency in adding like fractions but has difficulty with unlike fractions, the system detects this specific knowledge gap. It then assigns practice tasks that directly target the identified area of difficulty. These tasks are presented individually within the domain labeled *knowledge gaps*, enabling focused and tailored practice. This mechanism corresponds to the outer loop according to VanLehn. The structure of this system also qualifies Bettermarks as an ITS, in line with the definition of Ma et al. (2014) mentioned in section 2.2.

There were no guidelines for the teachers on how the ITS should be used. Teachers and students could decide freely on the extent of ITS usage and when it occurred (whether during or outside school hours). There were also no guidelines regarding the topics or usage format (group or individual work) although the design of the ITS lends itself more to individual use.

4.4 Instruments

4.4.1 Performance test

The performance test was constructed based on the content areas of the curriculum for grades 7 and 8 in the federal state Schleswig-Holstein (Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein, 2014). The aim was to develop a curriculum-valid measure of students' mathematical performance. Since the exact sequence in which these curriculum's topics are taught is determined by individual schools, it was not feasible to anchor specific content areas to a particular grade or to develop separate tests for grades 7 and 8. Instead, using the curriculum as a basis, a comprehensive test for grades 7 and 8 was developed. Content areas that require more

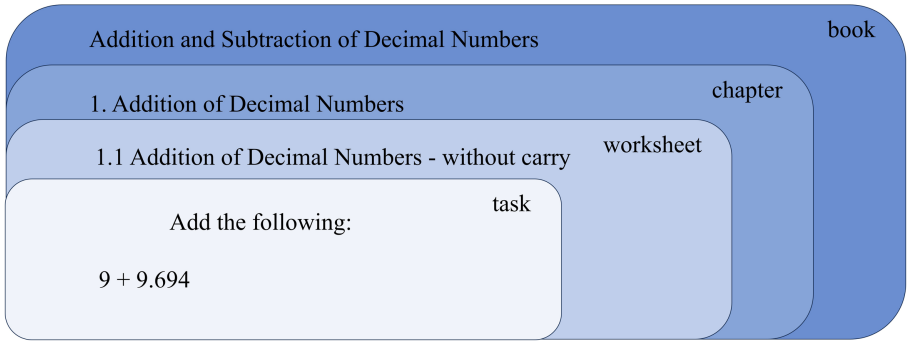

FIGURE 1
Exemplary content structure of bettermarks.

advanced background knowledge, such as logarithms and trigonometric functions were excluded, as they were unlikely to be taught before grade 9. The resulting test covered 12 content areas. The first four content areas—decimal numbers, common fractions, percentages/interest calculations, and variables/expressions—were each assessed using six items. The subsequent six areas—negative numbers, area calculations for polygons, linear equations, representations of functions, proportional relationships/rule of three, and linear functions—were each represented by five items. The final two areas—linear equation systems and inverse proportional relationships—were each covered by three items. To construct an appropriate item pool, curriculum-valid items from the International Trends in Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP) for 8th grade and the Program for International Student Assessment (PISA) study were selected. These items have been used in several studies over decades and have been optimized based on theoretical concepts and empirical evaluations (Martin and Kelly, 1996). Therefore, they provide a recognized and internally valid means of measuring mathematical performance and possess good test-theoretical properties. Among these curriculum-valid items with good test-theoretical properties, items that primarily assessed procedural knowledge (Hiebert and Lefevre, 1986) were selected. This decision was motivated by the intention to align the focus of the performance test with the types of tasks provided by the ITS, which primarily support the practice and consolidation of procedural knowledge.

To cover all content areas appropriately, four additional items had to be designed. Thus, the item pool consisted of 60 items in total (48 TIMSS items, seven NAEP items, one PISA item, and four self-constructed items). Of these, 55 items were single-choice items, and five were constructed-response items.

The items were administered in a multi-matrix design (Youden design), with the 60 items divided into six clusters of 10 items each. Each cluster contained items from 10 of the 12 content areas and each performance test consisted of two clusters, i.e., every student answered 20 items.

4.4.2 Noncognitive measurements

Besides the mathematics performance as a cognitive measure, several noncognitive measures were administered. The influence of constructs such as *subject-specific self-concept*, *mathematics anxiety*, *cost-utility*, and *work ethics* on student mathematics performance is empirically well-established (Hattie, 2023). Therefore, these constructs were captured in the questionnaire to be used as covariates in the analysis. We also decided to capture the constructs of *subject-specific self-concept* and *subject interest* in the subjects of German and English to use this information in the background model (see section 3.6.4). Additionally, to capture the students' experience using the ITS, they were asked to assess the ITS in terms of *usefulness*, *demand*, *affective and cognitive engagement*. The constructs were assessed using Likert scales, with each Likert scale comprising three to nine items on a four-point scale (*strongly disagree*, *disagree*, *agree*, *strongly agree*). The internal consistency of the administered noncognitive scales (Cronbach's α) ranged from 0.79 to 0.94.

In addition, the questionnaire asked for some personal data, including the cultural capital of the household (books-at-home-question), as this also has an empirically proven influence on student academic performance (OECD, 2023), as well as age, gender, school type, and semester grades. Finally, the students were asked to evaluate the ITS by grading its perceived usefulness on the standard German grading scale, ranging from 1 (*very good*) to 6 (*insufficient*).

4.5 Operationalization ITS usage

To investigate the impact of the frequency of ITS usage on learning progress, we first needed to operationalize what "usage" means in a way that fits the structure of the available data. The log data contained one entry for each time a student opened a worksheet, including the start date and time, but they do not contain any information on when students stopped working. Because of this limitation, we were not able to calculate actual time-on-task and instead focused on how often students initiated work on worksheets. To capture different facets of students' engagement with the ITS, we developed five plausible operationalizations of usage frequency.

The five operationalizations are defined as follows:

1 Number of worksheets opened.

This operationalization counts every instance of a student opening a worksheet during the measurement period, including repeated interactions with the same worksheet. It is conceptually close to common indicators of engagement, such as time spent using the system, as it reflects the overall volume of interactions. However, it may overrepresent short, dense bursts of activity (e.g., repeated re-openings), which can inflate usage counts without necessarily reflecting meaningful learning activity.

2 Number of unique worksheets opened.

Here, each worksheet is counted only once, regardless of how often it is revisited. This measure emphasizes the breadth of content students engaged with and reduces inflation due to repeated openings of the same worksheet. At the same time, it ignores repetition and revisiting, which may be pedagogically relevant, and is less closely related to time-based measures of system use.

3 Number of unique worksheets opened per hour.

In this approach, each worksheet is counted at most once per hour. This reduces the influence of very dense activity within short time spans while still allowing repeated engagement with the same worksheet to be reflected over time. A limitation is that the choice of an hourly window is somewhat arbitrary, and the measure may overestimate the usefulness of engagement when the same worksheets are worked on repeatedly.

4 Number of unique worksheets opened per day.

This operationalization counts each worksheet at most once per day. It further reduces the impact of short-term bursts of activity and aligns well with daily study routines and typical homework structures. However, it masks variation in within-day intensity and duration of work and treats brief and extensive engagement on the same day equivalently, resulting in a weaker correspondence with time-based engagement measures.

5 Days on which the ITS was used.

This measure captures the number of days on which at least one worksheet was opened. It emphasizes the regularity of engagement over time and is closely related to concepts of spaced practice. At the same time, it provides no information about the amount or depth of work completed on a given day and does not distinguish between minimal and extensive daily usage.

Because the period between pre- and posttest differed across classes, all operationalizations were adjusted by dividing the absolute count n by the number of weeks t between the two measurements. This yields the standardized usage score.

$$\text{ITSU} = \frac{n}{t}$$

We consider all five operationalizations valid, as each captures a different facet of how students used the ITS. Since none is theoretically superior, we selected the operationalization with the strongest raw effect size for the main analyses. Raw effect size refers to the association between usage and learning gains without adjusting for covariates. Section 5.1 reports the comparison of these raw effects and identifies which operationalization was chosen for further analysis.

4.6 Data analysis

4.6.1 Handling of missing data

Some of the questionnaires and tests contained missing data. Depending on the type of data, we took different approaches. A missing answer in the performance test was scored with zero points, since this meant that a student was either unsure of the answer and skipped the question or that the student was too slow to reach this question. Missing data in the covariates, however, was imputed not to lose students with only a few missing entries. Since for any given variable, only a small part of the data were missing (<5% in most cases), the missing data of the covariates was imputed via the R Package *missForest* (v1.5) (Stekhoven, 2022).

4.6.2 Descriptive statistics

The items were dichotomously rated as correct or incorrect. The solution rate in the pretest was 46%, which improved to 54% in the posttest. The data from pre- and posttest were scaled by a Rasch model that allowed us to score the students' performance on the logit metric. The average learning gain was approximately 0.39 logit over 250 days, which extrapolates to a learning gain of 0.58 logit per schoolyear. Studies on mathematics performance in Germany with similar students found average increases from 0.5 SD (PALMA) (vom Hofe et al., 2009) to 0.6–0.7 SD per year (Köller et al., 2000), which fit these results very well, assuming a sample standard distribution of one logit.

4.6.3 Quality assessment

4.6.3.1 DIF analysis

To assess the reliability of the performance tests, a graphical DIF analysis was conducted between the pretest and posttest groups (Figure 2).

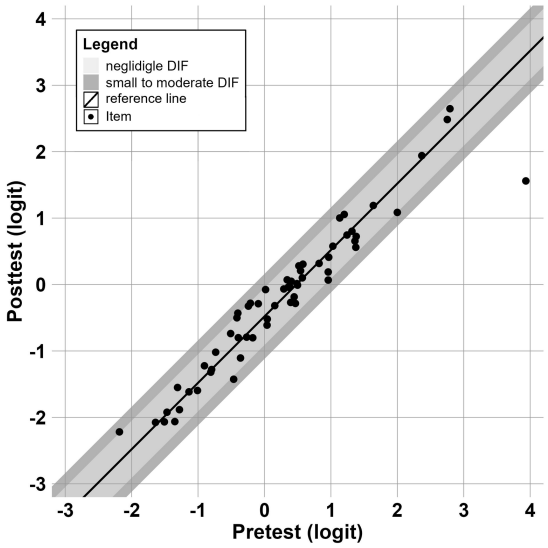

FIGURE 2
Graphical DIF analysis comparing pretest and posttest item difficulties.

The item difficulties were estimated separately for both tests and the resulting difference in difficulty for each item was compared. Ideally, the increase in mathematics performance should result in a uniform decrease in item difficulty for all items (better performing students solve items more frequently and therefore, by definition, the item difficulty decreases).

If the difficulty change of an item significantly deviates from this trend, it is referred to as *differential item functioning* (DIF) (Zumbo, 1999), as the items function (i.e., what is measured by that item) is different in both groups.

Tristán (2006) considers an absolute DIF of less than 0.43 logits as negligible, between 0.43 and 0.64 logits as small to moderate, and greater than 0.64 logits as moderate to large. According to these rules of thumb, four items show small to moderate DIF, and one item shows moderate to large DIF. These items were not excluded directly, but further quality assessment was conducted in the form of infit and item-total correlations.

4.6.3.2 Item quality

To assess the quality of the Rasch model, item-total correlations and infits were computed. The item-total correlation is a classical measurement to access the discrimination of an item. The item-total correlation for the utilized items ranged between 0.24 and 0.59, indicating an acceptable strength of discrimination.

The weighted mean square (infit) is a standard measure for assessing the fit of a Rasch model. In an item response theory (IRT) model, a specific level of variance is expected for each item. An infit value of 1.0 indicates the presence of the expected level of variance. A value above 1.0 suggests more variance (thus less discrimination) than expected, while a value below 1.0 suggests less variance (thus more discrimination) than expected. Acceptable values rage from 0.8. to 1.2 (Harks et al., 2014). The infit values ranged between 0.87 and 1.15, indicating a good fit to the model. As all items demonstrated satisfactory infit statistics,

they were retained for further analysis—even the one item that exhibited moderate to large differential item functioning (DIF).

4.6.4 Model structure and plausible values

To use all available information, item difficulties and person scores were estimated in a two-step-process using virtual persons (Tabachnick and Fidell, 2019). In the first step, all 2,982 tests were treated as if submitted during a single measurement point instead of being submitted from pre- and posttest. The item difficulties were then estimated with a unidimensional Rasch model according to Hartig and Kühnbach (2006), based on the result of all of these (partly virtual) students. In the second step, these item difficulties were fixed and used to estimate the person abilities of the 940 students in the longitudinal sample. The estimation of the item difficulties was conducted with an EAP reliability of 0.83 and a WLE reliability of 0.76, indicating good precision (Kline, 2000).

To generate plausible values (PVs), an extensive background model was applied, which accounted for the hierarchical structure of the dataset. When analyzing samples as a whole with Rasch models, PVs are often preferred over WLEs due to their superior performance (Mislevy et al., 1992). Because of the uncertainty inherent to IRT, each student is assigned both an estimated ability score as well as the variance of that score as a measure of uncertainty. The estimated score and the corresponding variance together with the variables in the background model are used to create a probability distribution of possible person ability scores for each student (posterior distribution). PVs are random draws of the ability score based on this posterior distribution.

The PVs for this analysis were generated according to Hartig and Kühnbach (2006), using a two-dimensional Rasch model with pre- and posttest forming the two dimensions. Fifty PVs were drawn for each pre- and posttest of the 940 students in our longitudinal study.

To model the structure of the data (students within classes within schools) adequately, we employed hierarchical linear modelling (HLM). HLM represents a method to analyze effects both on an individual level and on a class level while reflecting and preserving the inherent structure of the dataset (Field et al., 2012). In this case, level 1 represents individual students, while level 2 represents the different classes. Since the classes were from 13 different schools, adding a third level to reflect this structure in the model would have been possible. We decided against a third level partly because there was very little variance at the school level (ca. 5%) and partly because there were not enough schools for reliable parameter estimation.

5 Results

We begin by presenting the descriptive analysis of ITS usage, followed by the results from students' perspective on the ITS. The model development and the resulting multilevel models will be reported last.

5.1 ITS usage

Overall, the students completed 61,051 worksheets, indicating that each student on average completed about 2 worksheets per school week ($SD = 2.4$). The average solution rate was 34%, meaning roughly one in three given answers was correct. Knowledge gaps (see section 3.5) comprised 1.5% of the opened worksheets. If students worked on worksheets multiple times, each instance was counted separately. Approximately one-third of ITS usage occurred during school hours (Monday to Friday, 8 a.m. to 2 p.m.), while the remaining two-thirds occurred at home.

The operationalizations for ITS usage discussed in 4.5 were compared in terms of their raw effect size on both the class level (class mean) and the individual level (centered at the respective class means). The standardized effect sizes are shown in Table 1.

As can be seen in Table 1, operationalization 2 (counting each worksheet only once per student, no matter how often it was completed during the school year) has the strongest significant raw effect size on the posttest of all operationalizations. Therefore, this operationalization was used for subsequent analysis. Under this operationalization, the ITS usage is visualized in the histogram in Figure 3.

The average ITS usage is about 1.3 *unique* worksheets per school week. A total of 469 students completed less than one worksheet per school week, 301 completed between one and two, 170 completed 2 or more. It can be seen that the ITS usage varied between students and the ITS usage itself is substantial. Therefore, a potential effect of ITS usage on learning gains should show up in the final model.

5.2 Students' perspective on the ITS

The items assessing students' perspective on working with the ITS—specifically in terms of demand, affective and cognitive engagement, and usefulness—were averaged into their respective Likert scales.

As shown in Figure 4, most students found working with the ITS to be relatively low in demand ($M = 1.9$, $SD = 0.6$). They reported moderate levels of engagement ($M = 2.5$, $SD = 0.8$), and half of the

TABLE 1 Raw effect sizes for different operationalizations of ITS usage.

Operationalization	Raw size (without pretest)		Raw size (with pretest)	
	Class level	Individual	Class level	Individual
1 (All worksheets)	0.03	0.06	0.00	0.03
2 (Unique worksheets)	−0.01	0.22*	−0.02	0.09*
3 (Unique worksheets per hour)	−0.02-	0.19*	−0.02	0.08*
4 (Unique worksheets per day)	−0.02	0.15*	−0.02	0.07
5 (Days used)	−0.01	0.06	0.00	0.05

*$p < 0.05$.

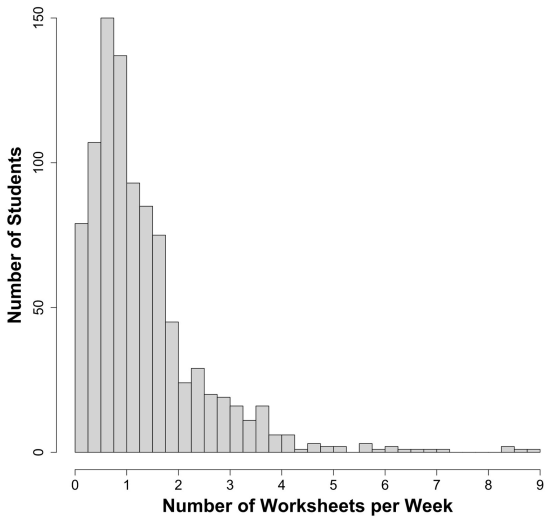

FIGURE 3
Distribution of weekly unique worksheet usage in the ITS.

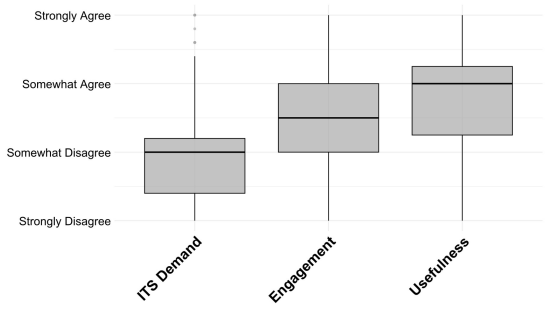

FIGURE 4
Distribution of student ratings for ITS demand, engagement, and usefulness.

students somewhat agreed that the ITS was helpful in learning mathematics ($M = 2.8$, $SD = 0.7$).

On the German grading scale (ranging from 1, *very good,* to 6, *insufficient*), students rated the ITS perceived usefulness with an average score of 2.7 ($SD = 1.0$). Overall, most students expressed neutral or mildly positive feelings about using the ITS on both the subjective ratings and the grade.

5.3 Model development and results

As described in the Methods section, a series of multilevel models (Table 2) were used to analyze the data. To build a model with a good fit to the dataset, the initial model was progressively made more complex by successively adding predictors. Predictors were added generally following the approach outlined by Field et al. (2012), if (a) there is empirical evidence which links the predictor to student performance and (b) adding the predictor increased the predictive power of the model. Models were compared based on their Akaike information criterion (AIC) and Bayesian information criterion (BIC) values. Both information criteria are measures of goodness of fit. While the values cannot be interpreted in an absolute way, a lower value indicates a better fit. In model M0, only the hierarchical structure of data with no predictors was considered. The intra class correlation (ICC) was 0.53, indicating there was about as much variance on the individual level as on the class level. In models M1 and M2, assessments of prior knowledge (*pretest score* and *marks in mathematics*) were added. In model M3, the noncognitive measures regarding mathematics and students' perspective were added. The five different noncognitive measurements regarding mathematics were expectedly moderately correlated (on average 0.43), so it made sense to only add one of them as a covariate. Out of five different scales, attitude towards mathematics had the strongest effect of the posttest score and was therefore included in model M3. In model M4, structural variables at the student and school level—grade and school type—were added. School type showed a substantial association with posttest performance, even after controlling for prior achievement and other covariates. Grade level, in contrast, was included primarily as a theoretically relevant control variable to account for differences in curricular progression, but did not exhibit an additional effect once prior performance was taken into account. As shown in Table 2, the stepwise addition of predictors from models M1 to M4 was associated with successive improvements in model fit, as indicated by decreasing AIC and BIC values, and with increases in explained variance at both the individual and class level. Including any other predictors, such as gender or self-concept in other subjects, resulted in an increase in both information criteria, AIC and BIC.

To examine the influence of ITS usage frequency, this factor was incorporated into model M5 at two levels. At the class level, the group-centered mean of ITS usage frequency was introduced for each class. At the individual level, the usage frequency of individual students, centered on the class mean, was included as part of the model. At the class level, we investigated whether classes that used the ITS more frequently (and thus probably traditional materials less frequently) showed greater learning gains compared to classes with lower usage frequency (between-group effect). At the individual level, we examined whether students who completed more tasks than their classmates exhibited higher learning gains (within-group effect). A between-group effect could therefore be interpreted as an improvement in learning gains compared to traditional learning materials, while a within-group effect could suggest that increased learning time enhances learning gains.

Table 2 presents the hierarchical inclusion of predictors across models M1 to M5. In M1, performance in the pretest is a strong predictor of the performance in the posttest, accounting for approximately 56% of the variance at the individual level and approximately 83% at the class level. The inclusion of marks as an additional predictor in M2 increases the proportion of explained variance and improves model fit. Model M3 shows that a more positive attitude towards mathematics is significantly associated with higher posttest scores. In M4, the influence of school type becomes evident. Controlling for prior performance, mathematics attitude, marks in mathematics and grade, the difference in posttest performance between students attending academic track schools (*Gymnasium*) and those from comprehensive schools (*Gemeinschaftsschule*) is 0.57 logits, indicating substantially greater learning gains among students in the academic track. The inclusion of ITS usage as a predictor in M5 does not improve model fit, as indicated by higher AIC and BIC values compared to Model 4, and the explained variance remains largely unchanged. Furthermore, ITS usage does not

TABLE 2 Effect sizes and information criteria of different multilevel models.

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Level 1 (students)					
Pretest	0.83*(0.03)	0.77*(0.04)	0.77*(0.04)	0.75*(0.04)	0.75*(0.04)
Marks[a]		0.12*(0.03)	0.11*(0.03)	0.11*(0.03)	0.11*(0.03)
Math attitude[b]			0.06*(0.03)	0.06*(0.03)	0.06*(0.03)
ITS usage[b,c]					0.06 (0.03)
Level 2 (classes)					
School type				0.57*(0.10)	0.58*(0.10)
Grade				0.17 (0.09)	0.17 (0.09)
ITS-usage (class mean)[b]					−0.01 (0.05)
Explained variance					
Individual level	55.9%	57.8%	58.3%	58.4%	58.5%
Class level	83.5%	81.0%	80.6%	91.0%	90.9%
			Model fit		
AIC	1616.41	1538.27	1536.14	1509.08	1517.54
BIC	1635.80	1562.32	1565.01	1547.57	1565.66

Unless otherwise noted, predictors were not standardized. The standard error is given in brackets. ITS, intelligent tutoring system. [a]in Germany, lower marks correspond to better performance, therefore the scale was inverted. [b]Predictor was z standardized. [c]Predictor was group-mean centered. *$p < 0.05$.

exhibit a significant effect on student performance at either the individual or class level.

6 Discussion

6.1 Conclusion

Based on a longitudinal pre–posttest design, our study examined whether the frequency of ITS use influenced students' learning gains. The multilevel analysis revealed a significant positive effect of ITS usage on mathematics performance ($\beta = 0.22$) on the individual level when not controlling for other variables (left part of Table 1). However, when controlling pretest scores, the observed effect of ITS usage on mathematics performance was notably smaller ($\beta = 0.09$, right part of Table 1), a pattern consistent with findings from a meta-analysis by Ma et al. (2014), which showed that studies accounting for baseline differences tend to report smaller effect sizes. This smaller, yet still significant effect aligns with results reported by Spitzer (2022) for the same ITS; however, that study did not account for additional covariates beyond prior performance. In contrast, in our study, when further covariates—such as students' attitudes toward mathematics, grade level, and school type—were included in the model, the effect of ITS use was no longer statistically significant, neither on the individual nor the class level (Model 5 in Table 2). Therefore, the answer to the research question is that no significant relationship was found between the frequency of ITS use and student learning gains in curricular mathematics tasks, once prior performance and other relevant covariates were controlled for. This non-significant effect shows the importance of controlling relevant covariates in a correlational design, even though doing so substantially increases the analytical workload.

The analysis was repeated using alternative operationalizations of ITS usage (as defined in section 4.5), including counting every opened worksheet, limiting each worksheet to one count per hour or per day, and counting only days in which the ITS was used at least once, yielding no significant deviations from the aforementioned results. The effect of ITS use on posttest performance was smaller once pretest performance was controlled for. This indicates that high-achieving students use the ITS more and learn more over the course of a school year. However, our findings do not support the assumption that increased ITS usage directly leads to greater learning gains. Simply using the ITS more frequently—potentially at the expense of other instructional approaches—does not appear to enhance student learning outcomes. As Higgins et al. (2012) aptly put it, "…it is not whether technology is used (or not) which makes the difference, but how well the technology is used to support teaching and learning" (p. 3). In addition, students perceived the ITS as having a moderately positive impact on their mathematics learning. This aligns with the findings from the multilevel model, which suggest that the ITS supports mathematics learning to a similar extent as traditional media.

6.2 Limitations

As the study was conducted in the context of the introduction of the ITS in a federal state in Germany, it was not possible to conduct an experiment with random assignment to a control group not using the ITS. Therefore, the lack of a control group restricts our ability to firmly establish causal effects of the ITS's impact on learning outcomes. Nevertheless, we found substantial variance in the "natural" usage of ITS among different classes and we could show a large discrepancy between raw and controlled effect sizes of ITS usage, both of which will be important for further investigation.

Due to the limited available log data, it was only possible to analyze the usage frequency of the students. This reduces the explanatory power of our analysis, as other factors—such as the social

form employed during ITS usage as well as the way teachers implemented the ITS during and beyond school hours—could also influence the ITS's effectiveness.

As an additional consequence resulting from the limited log data, ITS usage was operationalized as the frequency with which students opened unique worksheets. As a result, it was not possible to distinguish between different forms of engagement, such as careful, effortful work and rapid task completion, nor to identify non-learning behaviors. This limitation may have weakened the observed relationship between ITS usage and learning gains. Nevertheless, the moderate raw association between ITS usage and performance suggests that this measure still captured a meaningful, though coarse, indicator of students' interaction with the ITS.

6.3 Reasons for the lack of an ITS effect

The results of our study raise the question of why the use of the ITS was not effective. In the following, several plausible reasons are discussed. Because the extent to which students actually used the system determines the potential impact of any instructional technology, the discussion first considers the intensity of ITS use observed in our data. After outlining the extent of actual system use, the discussion begins with specific features of the ITS itself including the types of tasks available and the nature of the feedback, and the extent to which the system's adaptive functionality was utilized. Then moves on to challenges related to its use in the classroom, including self-pacing, homework integration, collaborative learning, and the role of teachers in shaping instructional use. Furthermore, a broader perspective is taken by examining the ITS's alignment with principles from CTML and guided activity, before reflecting on aspects of the study design—such as the long measurement period—that may also have played a role. The section concludes with a discussion of the alignment between the ITS and the performance test.

Although students completed a total of 61,051 worksheets, the average individual usage corresponded to only about 1.3 unique worksheets per school week, with nearly half of the students completing fewer than one unique worksheet per week. Taken together with the null effects in our multilevel models, this suggests that many students may not have reached an intensity of ITS use that is sufficient to produce detectable gains on a broad curriculum-based test.

One contributing factor may be the limited types of tasks that students could engage with in the ITS. Although the system offers a large number of items, many of them focus on highly procedural exercise formats and comparatively few engage students in forms of mathematical activity linked to deeper learning, such as constructive reasoning, modelling real-world situations, communicating mathematical ideas, sketching representations, or working through open or non-routine problems. These forms of activity are widely recognized as essential components of mathematical proficiency because they require students to make sense of mathematical structures, connect representations, and articulate their reasoning (e.g., Santos-Trigo, 2024). As outlined in the introduction, the ITS primarily offers tasks with clearly structured scaffolding and step-by-step guidance, but includes relatively few opportunities for tasks that actively promote such constructive or representational reasoning. In this regard, the system may be well suited for supporting procedural fluency but less effective for fostering flexible and conceptual mathematical understanding.

The type of feedback provided by the ITS could also account for the results. Although the system offers immediate feedback at each substep, the depth and quality of this feedback vary considerably: some items provide elaborated explanations, whereas many others offer only corrective "right/wrong" responses. Research has shown that elaborative feedback is generally more beneficial for learning than purely corrective feedback because it supports the development of conceptual understanding and helps students diagnose their misconceptions (Van Der Kleij et al., 2015). In our study, this pattern aligns with students' own perceptions: they rated the ITS as only moderately engaging and only somewhat helpful for learning, suggesting that the predominantly corrective feedback may have facilitated procedural success without consistently fostering deeper understanding. For more complex or conceptually demanding tasks, students may therefore have required additional forms of scaffolding beyond immediate correctness indicators.

Another factor that may help explain the absence of significant effects concerns the limited use of the ITS's adaptive knowledge gap functionality. Although the system is designed to identify and address individual misconceptions through targeted follow-up tasks, knowledge gaps accounted for only 1.5% of all opened worksheets, despite an average solution rate of approximately one third (section 5.1). This indicates that, while students frequently made errors during regular worksheet work, these errors translated into only minimal engagement with targeted follow-up tasks addressing identified difficulties. One possible explanation is that knowledge gaps were assigned relatively conservatively by the system; another is that students often did not engage with recommended follow-up tasks once they were available. The present data do not allow a clear distinction between these explanations. Previous research suggests that learning effects of digital tools are particularly pronounced when systems are adaptive (Hillmayr et al., 2020). Against this background, the largely underutilized adaptive functionality observed in the present study provides a plausible explanation for why the ITS did not yield stronger effects.

In this context, how teachers handled the use of the ITS also likely influenced the effectiveness. For the portion of ITS usage that occurred during school hours, it is not entirely clear to what extent students were able to work with the ITS at their own pace, as this likely depended on how much autonomy teachers allowed during ITS use. As found by Moreno (2007), allowing students to self-pace lowers perceived difficulty.

In addition to usage during school hours, a substantial part (approximately two thirds) of the ITS usage was conducted outside school hours. This indicates that a significant fraction of homework was completed and handed in through the ITS. As found by Hattie (2023), the evaluation and discussion of homework is important for students. If the digital homework was not properly discussed in class because the ITS provided worked examples, this could hinder students to ask questions and get clarification about certain tasks and therefore negatively impact their learning gains.

Another important factor is collaborative learning, which has been shown to enhance learning outcomes. As outlined in the theoretical framework, students who engage in collaborative learning tend to learn more effectively, especially when technology facilitates this process (Chen et al., 2018; Sung et al., 2017). However, the ITS

analyzed here is not designed to support collaborative learning in pairs or small groups. Given that the majority of ITS use in our sample occurred outside school hours, much of the work with the system likely took place individually rather than in pairs or groups. This usage pattern, together with the lack of collaborative features in the system, may help explain why potential benefits of collaborative learning with technology did not materialize here. Consequently, greater use of the ITS may limit opportunities for students to collaborate and engage in meaningful discussions about mathematical concepts. This lack of interaction could result in smaller learning gains overall.

Teachers' decisions regarding the adoption of the ITS constitute another important factor in interpreting the findings. Although the system was available to all participating schools, its use was entirely voluntary, and teachers retained full discretion over whether they used the ITS at all and how intensively it was employed in their classes. Even though the ITS was adopted in many classrooms, overall usage intensity remained low and varied substantially across classes, indicating that mere availability did not translate into sustained or systematic use. This pattern is consistent with research showing that teachers' adoption of digital tools depends strongly on their perceived usefulness, attitudes toward technology, and institutional conditions (Teo, 2011). Moreover, when digital tools are not embedded in shared instructional routines at the school level, uptake tends to remain uneven, which is typically associated with reduced or absent effects on student learning (Ertmer and Ottenbreit-Leftwich, 2010).

Beyond adoption, the instructional use of the ITS is therefore likely to have influenced its effectiveness. Although teachers who chose to use the system were free to embed it into their teaching as they saw fit, no mandatory training or instructional guidance accompanied its implementation. Meta-analytic evidence indicates that digital interventions yield substantially larger learning effects when teacher training is provided than when it is not (Hillmayr et al., 2020). Teacher training typically focuses not only on technical operation but also on how digital tools can be integrated into instruction, how students' work can be monitored, and how digital tasks can be connected to classroom discussion and follow-up activities. As a result, the pedagogical use of the ITS likely varied widely, and in some cases may not have been closely aligned with instructional goals or curricular demands. Earlier experimental work further suggests that insufficient preparation can, in some cases, even be associated with negative learning effects (Koedinger and Anderson, 1993). Taken together, these findings highlight the teacher's role in shaping the effectiveness of digital tools and provide a plausible explanation for the null effects observed under conditions of voluntary and unsupported ITS use.

In addition to the lack of collaborative learning, the results can in part be explained with CTML (Mayer, 2014). Since most tasks found in the ITS operate at a symbolic or algebraic level, according to CTML, they do not make full use of the pictorial channel, leading to less effective learning.

In contrast to the previously discussed theoretical concepts, the ITS adheres to the principles of guided activity very well. The ITS incorporates well-structured substeps with clear instructions, aligning with the principles of guided activity. This design facilitates task completion by providing explicit guidance to students and, hence, assumingly fosters effective learning.

The study design itself also plays a role in interpreting the results. As Kulik and Fletcher (2016) found, longer measurement periods correlate with smaller effect sizes. One possible explanation for this is that the novelty factor of the ITS diminishes over time.

The ITS and the performance test were closely aligned with respect to both curricular content and the types of competencies addressed. As described in sections 4.3 and 4.4, both were developed with reference to the same state-level curricula, resulting in substantial overlap in the mathematical content covered. In addition, the performance test primarily assessed procedural knowledge, which corresponds to the predominant focus of the tasks provided by the ITS. The observed pre–post learning gains were of a magnitude that is typical for students at this grade level over a school year, suggesting that the performance test was sensitive to the mathematical learning that occurred during the measurement period. Against this background, it is unlikely that the absence of a significant effect of ITS use can be attributed to a mismatch between the learning environment and the performance test.

In summary, while the ITS aligns well with the principle of guided activity, other aspects may help explain its limited effectiveness. The types of tasks and the feedback provided may not have been sufficient to support deeper learning. Opportunities for self-paced work in class were likely restricted, and digital homework may have lacked in-class follow-up. Key principles from CTML also appear only partially addressed. Lastly, the long measurement period may reduce the measured effect sizes in digital learning due to diminishing novelty effects.

Considering limitations of the ITS and likely challenges in its integration into classroom settings, the absence of significant effects are reasonable. They align with Hattie (2023) findings that the mere presence of technology in the classroom does not guarantee improved learning outcomes. Instead, it is probably the way technology is integrated into the teaching process that truly matters.

Given the ecologically valid nature of the study, the results can be interpreted as a realistic evaluation of the implementation of an ITS. In Germany, if schools are provided with technology, such as an ITS, the frequency and manner in which this technology is used is almost entirely at the discretion of the individual teacher. The hope associated with the provision of an ITS is that this type of technology use will optimize learning processes and lead to sustainably higher performances in international (PISA) and national standardized evaluations in Germany.

6.4 Implications

A more precise delineation of the capabilities and limitations of digital media in education is necessary. In particular, it is important to address how effective digital tools are in supporting students when revisiting familiar concepts as opposed to engaging with new ones. Also, exploring the importance of students' knowledge about the ITS and their metacognitive awareness of learning with ITS merit attention.

6.4.1 Comprehensive analysis approach

Current research often focuses either on measuring the effectiveness of ITS with limited process data or on analyzing process data without adequately considering external factors such as performance in standardized tests and relevant noncognitive measurements. To gain a deeper understanding of why ITS can

facilitate effective learning, a holistic approach is needed. By combining outcome-based evaluation with contextual and usage-related information, the present study represents a substantial step in this direction. Access to even more comprehensive data—effectiveness measurement and process analysis— would allow for an even more nuanced analysis of how the ITS was used and how it impacts learning. Therefore, more comprehensive data on feedback received, utilization patterns, and engagement with features like worked examples would be useful for understanding how the ITS was used. In addition, data on student use of the ITS and teacher support inside and outside classes is also important for identifying efficient ITS usage. Gathering and analyzing all this information (learning gains and relevant noncognitive measurements, detailed log data, and process data provided by students and teachers) would possibly allow finding effective implementations of the ITS in teaching and learning.

In summary, future research should strive to adopt a broad perspective and comprehensive analysis approach, aiming to uncover general underlying trends rather than focusing solely on specific features of individual systems. Addressing these research desiderata will enhance our understanding of how digital tools can be optimally leveraged to support teaching and learning processes.

6.4.2 Potential of digital media in learning mathematics

Our findings suggest several additional directions for future research on the use of digital media in education. One area that deserves attention is collaborative learning. Collaboration—especially when supported by digital tools—can improve learning outcomes (Chen et al., 2018). However, many digital learning environments are primarily designed for individual use. This may be because it is difficult to combine personalized instruction with group-based learning. Still, finding ways to enable student interaction and exchange within digital learning settings could help increase their overall effectiveness.

Another important point concerns the type of feedback students receive. In many systems, feedback is limited to whether an answer is correct or not. But research shows that elaborated feedback—explaining why an answer is right or wrong—leads to better learning outcomes. New developments in artificial intelligence could make it easier to provide such feedback in a flexible and adaptive way. For example, AI-based systems could react to specific student errors without the need to pre-program all possible explanations, as a recent meta-analysis shows (Yi et al., 2025). Exploring this potential could be a valuable area for future research.

A third aspect relates to the use of multiple and dynamic representations. One of the key advantages of digital media is the ability to visualize abstract content in new ways—for instance, through interactive graphics or simulations. In mathematics education, examples include dynamic representations of fractions or functions. These tools offer possibilities that traditional media cannot, but more research is needed on how to integrate them effectively into teaching.

Although these points are based on the analysis of one ITS, they apply to digital learning tools more broadly. Similar challenges and opportunities can be found in other technologies such as *dynamic geometry environments* (DGEs) or *spreadsheets*. The findings support the idea that it is not the technology itself that determines its effectiveness, but how it is used in teaching. Simply introducing digital tools into the classroom is not enough—what matters is how they are embedded into instruction, how students interact with them, and how teachers guide their use. Future research should therefore focus not only on the tools themselves, but also on the pedagogical strategies and learning environments in which they are applied.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Ethics Committee of the IPN – Leibniz Institute for Science and Mathematics Education. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

JS: Writing – review & editing, Writing – original draft, Formal analysis, Software, Data curation, Validation, Methodology, Visualization. TR: Software, Formal analysis, Methodology, Project administration, Conceptualization, Investigation, Writing – review & editing, Validation, Supervision. GN: Conceptualization, Supervision, Methodology, Writing – review & editing. AH: Supervision, Writing – review & editing, Project administration, Methodology, Conceptualization.

Funding

The author(s) declared that financial support was not received for this work and/or its publication.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. Generative AI was used to improve grammar and style of the manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akram, H., and Abdelrady, A. H. (2023). Application of classpoint tool in reducing EFL learners' test anxiety: an empirical evidence from Saudi Arabia. *J. Comput. Educ.* 10, 529–547. doi: 10.1007/s40692-023-00265-z

Akram, H., and Abdelrady, A. H. (2025). Examining the role of ClassPoint tool in shaping EFL students' perceived E-learning experiences: a social cognitive theory perspective. *Acta Psychol.* 254:104775. doi: 10.1016/j.actpsy.2025.104775

Atkinson, R. C. (1968). Computerized instruction and the learning process. *Am. Psychol.* 23, 225–239. doi: 10.1037/h0020791

Belland, B. R., Walker, A. E., Kim, N. J., and Lefler, M. (2017). Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. *Rev. Educ. Res.* 87, 309–344. doi: 10.3102/0034654316670999

Bettermarks. (2025) Das adaptive Lernsystem für Mathematik [The adaptive learning system for mathematics]. Available online at: https://de.bettermarks.com (Accessed April 3, 2024)

Chen, J., Wang, M., Kirschner, P. A., and Tsai, C.-C. (2018). The role of collaboration, computer use, learning environments, and supporting strategies in CSCL: a meta-analysis. *Rev. Educ. Res.* 88, 799–843. doi: 10.3102/0034654318791584

Ertmer, P. A., and Ottenbreit-Leftwich, A. T. (2010). Teacher Technology Change: How Knowledge, Confidence, Beliefs, and Culture Intersect. *Journal of Research on Technology in Education* 42: 255–84. doi: 10.1080/15391523.2010.10782551

Field, A., Miles, J., and Field, Z. (2012). Discovering statistics using R. Repr: Sage.

Harks, B., Klieme, E., Hartig, J., and Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educ. Assess.* 19, 243–66. doi: 10.25656/01:17987

Hartig, J., and Kühnbach, O. (2006). "Schätzung von Veränderung mit 'Plausible Values' in multidimensionalen Raschmodellen [Estimating Change with 'Plausible Values' in Multidimensional Rasch Models]" in Veränderungsmessung und Längsschnittstudien in der empirischen Erziehungswissenschaft. eds. A. Ittel and H. Merkens (Wiesbaden: VS Verlag für Sozialwissenschaften), 27–44.

Hattie, J. (2023). Visible learning: the sequel: a synthesis of over 2,100 Meta-analyses relating to achievement. 1st Edn. London and New York: Routledge.

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Hiebert, J., and Lefevre, P. (1986). "Conceptual and procedural knowledge in mathematics: an introductory analysis" in Conceptual and procedural knowledge: the case of mathematics. ed. J. Hiebert (Hillsdale, NJ: Lawrence Erlbaum Associates), 1–27.

Higgins, S., Xiao, Z. M., and Katsipataki, M. (2012). The impact of digital technology on learning: a summary for the education endowment foundation. Full report. Durham: School of Education, Durham University.

Hillmayr, D., Ziernwald, L., Reinhold, F., Hofer, S. I., and Reiss, K. M. (2020). The potential of digital tools to enhance mathematics and science learning in secondary schools: a context-specific meta-analysis. *Comput. Educ.* 153:103897. doi: 10.1016/j.compedu.2020.103897

Kline, P. (2000). Handbook of psychological testing. 2nd Edn. London and New York: Routledge.

Koedinger, K. R., and Anderson, J. R. (1993). "Effective use of intelligent software in high school math classrooms" in Proceedings of AI-ED 93: World Conference on Artificial Intelligence in Education. eds. P. Brna, S. Ohlsson and H. Pain (Charlottesville, VA: Association for the Advancement of Computing in Education (AACE)), 241–248.

Köller, O., Baumert, J., and Schnabel, K. (2000). "Zum Zusammenspiel von schulischem Interesse und Lernen im Fach Mathematik: Längsschnittanalysen in den Sekundarstufen I und II" in Interesse und Lernmotivation: Untersuchungen zu Entwicklung, Förderung und Wirkung. eds. U. Schiefele and K.-P. Wild (Münster: Waxmann), 163–181.

Kulik, J. A., and Fletcher, J. D. (2016). Effectiveness of intelligent tutoring systems: a meta-analytic review. *Rev. Educ. Res.* 86, 42–78. doi: 10.3102/0034654315581420

Kulik, J. A., and Kulik, C.-L. C. (1988). Timing of feedback and verbal learning. *Rev. Educ. Res.* 58, 79–97. doi: 10.3102/00346543058001079

Martin, M. O., and Kelly, D. L. (1996). Technical report. Chestnut Hill, MA: Center for the Study of Testing, Evaluation, and Educational Policy, Boston College.

MatheGym. (2025). Die Lernplattform für gymnasium und Realschule [The learning platform for academic track schools and comprehensive schools]. Available online at: https://www.mathegym.de (Accessed April 3, 2024).

Ma, W., Adesope, O. O., Nesbit, J. C., and Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: a meta-analysis. *J. Educ. Psychol.* 106, 901–918. doi: 10.1037/a0037123

Mayer, R. E. (2014). "Cognitive theory of multimedia learning" in The Cambridge handbook of multimedia learning. 2nd ed (Cambridge: Cambridge University Press).

Ministerium für Schule und Berufsbildung des Landes Schleswig-Holstein 2014 Fachanforderungen: Mathematik. Allgemein bildende Schulen – Sekundarstufe I – Sekundarstufe II [subject requirements: Mathematics. General education schools – Lower secondary level – Upper secondary level] Kiel Stamp Media & Schmidt & Klaunig Available online at: https://fachportal.lernnetz.de/sh/fachanforderungen/mathematik.html (Accessed April 1, 2021).

Mislevy, R. J., Beaton, A. E., Kaplan, B., and Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *J. Educ. Meas.* 29, 133–161. doi: 10.1111/j.1745-3984.1992.tb00371.x

Moreno, R. (2007). Optimising learning from animations by minimising cognitive load: cognitive and affective consequences of signalling and segmentation methods. *Appl. Cogn. Psychol.* 21, 765–781. doi: 10.1002/acp.1348

Moreno, R., and Mayer, R. (2007). Interactive multimodal learning environments: special issue on interactive learning environments: contemporary issues and trends. *Educ. Psychol. Rev.* 19, 309–326. doi: 10.1007/s10648-007-9047-2

Mullis, I. V. S., Martin, M. O., and von Davier, M. 2021. TIMSS 2023 assessment frameworks. Chestnut Hill, MA: International Association for the Evaluation of Educational Achievement. Available online at: http://www.iea.nl (Accessed December 12, 2025).

OECD (2023). PISA 2022 results (volume I): the state of learning and equity in education. Paris: OECD Publishing.

Santos-Trigo, M. (2024). Problem solving in mathematics education: tracing its foundations and current research-practice trends. *ZDM Math. Educ.* 56, 211–222. doi: 10.1007/s11858-024-01578-8

Shute, V. J., and Zapata-Rivera, D. (2007). Adaptive technologies. *ETS Res. Rep. Ser.* 2007, i–34. doi: 10.1002/j.2333-8504.2007.tb02047.x

Spitzer, M. W. H. (2022). Just do it! Study time increases mathematical achievement scores for grade 4-10 students in a large longitudinal cross-country study. *Eur. J. Psychol. Educ.* 37, 39–53. doi: 10.1007/s10212-021-00546-0

Steenbergen-Hu, S., and Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on K–12 students' mathematical learning. *J. Educ. Psychol.* 105, 970–987. doi: 10.1037/a0032447

Steenbergen-Hu, S., and Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *J. Educ. Psychol.* 106, 331–347. doi: 10.1037/a0034752

Stekhoven, D. J. 2022. missForest: nonparametric missing value imputation using random Forest. Version 1.5. R package. Comprehensive R archive network (CRAN). Available online at: https://CRAN.R-project.org/package=missForest (Accessed May 3, 2023).

Sung, Y.-T., Yang, J.-M., and Lee, H.-Y. (2017). The effects of Mobile-computer-supported collaborative learning: meta-analysis and critical synthesis. *Rev. Educ. Res.* 87, 768–805. doi: 10.3102/0034654317704307

Tabachnick, B. G., and Fidell, L. S. (2019). Using multivariate statistics. 7th Edn. Boston, MA: Pearson.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., and Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: a second-order meta-analysis and validation study. *Rev. Educ. Res.* 81, 4–28. doi: 10.3102/0034654310393361

Teo, T. 2011. "Factors Influencing Teachers' Intention to Use Technology: Model Development and Test." *Computers & Education* 57:2432–40. doi: 10.1016/j.compedu.2011.06.008

Tristán, A.. 2006. "An adjustment for sample size in DIF analysis." In Rasch measurement: Transactions of the Rasch measurement SIG, American Educational Research Association, edited by J. Bernoulli, W. P. Fischer, C. Shannon and G. Rasch, 20, 1070–1071. Available online at: http://www.rasch.org/rmt/rmt203e.htm (Accessed February 12, 2023).

Tullis, J. G., and Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *J. Mem. Lang.* 64, 109–118. doi: 10.1016/j.jml.2010.11.002

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse (2009). Cognitive tutor® algebra I: Intervention report (middle school math). Washington, DC: U.S. Department of Education.

Van Der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

VanLehn, K. (2006). The behavior of tutoring systems. *Int. J. Artif. Intell. Educ.* 16, 227–265. doi: 10.3233/IRG-2006-16(3)02

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educ. Psychol.* 46, 197–221. doi: 10.1080/00461520.2011.611369

vom Hofe, R., Hafner, T., Blum, W., and Pekrun, R. (2009). ""Die Entwicklung mathematischer Kompetenzen in der Sekundarstufe – Ergebnisse der Längsschnittstudie PALMA" [the development of mathematical competencies in secondary education – results of the longitudinal study PALMA]" in Mathematiklernen vom Kindergarten bis zum Studium: Kontinuität und Kohärenz als Herausforderung für den Mathematikunterricht. eds. A. Heinze and M. Grüßing (Münster: Waxmann), 125–146.

Yi, L., Liu, D., Jiang, T., and Xian, Y. (2025). The effectiveness of AI on K-12 students' mathematics learning: a systematic review and meta-analysis. *Int. J. Sci. Math. Educ.* 23, 1105–1126. doi: 10.1007/s10763-024-10499-7

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores. Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defence.