Education and Information Technologies (2025) 30:6019–6060
https://doi.org/10.1007/s10639-024-13025-y

Lessons learned from designing an open-source automated feedback system for STEM education

Steffen Steinert[1] · Lars Krupp[2,3] · Karina E. Avila[4] · Anke S. Janssen[5] · Verena Ruf[1] · David Dzsotjan[1] · Christian De Schryver[6] · Jakob Karolus[2,3] · Stefan Ruzika[4] · Karen Joisten[5] · Paul Lukowicz[2,3] · Jochen Kuhn[1] · Norbert Wehn[6] · Stefan Küchemann[1]

Received: 10 August 2023 / Accepted: 27 August 2024 / Published online: 25 September 2024

Abstract

As distance learning becomes increasingly important and artificial intelligence tools continue to advance, automated systems for individual learning have attracted significant attention. However, the scarcity of open-source online tools that are capable of providing personalized feedback has restricted the widespread implementation of research-based feedback systems. In this work, we present RATsApp, an open-source automated feedback system (AFS) that incorporates research-based features such as formative feedback. The system focuses on core STEM competencies such as mathematical competence, representational competence, and data literacy. It also allows lecturers to monitor students' progress. RATsApp can be used at different levels of STEM education or research, as it allows the creation and customization of the educational content. We present a specific case of its implementation in higher education, where we report the results of a usability survey (N=64), using the technology acceptance model 2 (TAM2), to evaluate the user experience of undergraduate students. Our findings confirm the applicability of the TAM2 framework, revealing that factors such as the relevance to the course of study, output quality, and ease of use significantly influence the perceived usefulness. We also found a linear relation between the perceived usefulness and the intention to use, which in turn is a significant predictor of the frequency of use. Moreover, the formative feedback feature of RATsApp received positive feedback, indicating its potential as an educational tool. Furthermore, as an open-source platform, RATsApp encourages public contributions to its ongoing development, fostering a collaborative approach to improve educational tools.

Keywords Automated feedback system · Online learning platform · STEM competence estimation · Learning scaffolds · Learning dashboards · Technoethics

Steffen Steinert, Lars Krupp and Karina E. Avila contributed equally to this work

Extended author information available on the last page of the article

1 Introduction

The demand for online learning opportunities is rising due to enhanced accessibility and interconnection (Harasim, 2000; Anderson, 2003). Even before the onset of the COVID-19 pandemic, which triggered an increase in demand for online learning materials, a rising number of students had turned to online resources to supplement their learning (Chandra & Palvia, 2021; Muthuprasad et al., 2021). To address this demand, a variety of applications are being developed to support students in their educational journey. One such type of system is an Automated Feedback System (AFS), which offers students immediate and customized instructions or feedback.AFSs encompass a broad range of educational tools, including Intelligent Tutoring Systems (ITS), teaching platforms, or online feedback tools (Deeva et al., 2021; Steinert et al., 2023). These systems share a common goal: to provide feedback that facilitates students' learning and improvement. One type of feedback of AFSs is formative feedback (Kulhavy & Stock, 1989; Butler & Winne, 1995; Hattie & Timperley, 2007), which is intended to modify the students' thinking or behavior to enhance their learning (Shute, 2008).

There are numerous AFSs available for science, technology, engineering, and mathematics (STEM) education. Each system has its own level of quality, which includes aspects such as reliability, performance, ease of use, and the adequacy of customer support. While it is beyond the scope of this article to discuss all of them, interested readers can refer to existing reviews on the subject (Deeva et al., 2021; Mousavinasab et al., 2021). In their review of AFSs, Deeva et al. (2021) found a shortage of open-source and well-maintained systems. The term open-source refers to software where the source code is freely available for anyone to use, examine, modify, and improve. During the writing of this manuscript, a new open-source web-based adaptive tutoring system called OATutor was published (Pardos et al., 2023). It includes knowledge tracing, which employs algorithms to monitor and predict student performance over time, an A/B testing framework that compares different educational strategies to find the most effective one, and learning tools interoperability support to ensure seamless integration of diverse learning tools and resources. Another example of an open-source adaptive system is COLT (Myers & Chatlani, 2017), which is web-based and designed to improve coding literacy. However, it has fewer features compared to OATutor. Open-source tools, like those mentioned above, have freely available source code that enables educators and developers to foster innovation and customize solutions to meet diverse educational needs.

In today's data-driven world, where extensive datasets are gathered and examined across several fields, proficiency in mathematics emerges as one of the key competencies of the 21st century (OECD, 2018). Consequently, adaptive systems that enhance mathematical skills are particularly important, as this knowledge is essential to navigate through data effectively, derive precise insights, and execute informed decision-making (Blumenthal, 1961; Hudson & Liberman, 1982; Hake, 2002). This is especially crucial since many students struggle with developing mathematical proficiency and these difficulties can have a negative impact on their performance in other subjects (Ballard & Johnson, 2004; Pozo & Stull, 2006; Llamas et al., 2012). For instance, those with a weak foundation in mathematics may have difficulty understanding scientific concepts or solving problems in engineering disciplines. This highlights

the need for effective educational interventions aimed at facilitating the development of students' mathematical skills and at providing support to teachers in their mathematical teaching methods. Such interventions may include technology-enhanced learning tools such as e-learning platforms, to improve the overall educational process.

In this article, we introduce RATsApp (Rapid Assessment Tasks App), an AFS developed by an interdisciplinary team of computer scientists, engineers, ethicists, as well as mathematics and physics education researchers. RATsApp is an open-source web-based application that is designed to meet the growing need for accessible and customizable tools in the field of STEM education. RATsApp combines the theoretical framework of formative feedback (Kulhavy & Stock, 1989; Butler & Winne, 1995; Hattie & Timperley, 2007) – immediate, ongoing feedback provided during the learning process to modify students' thinking or behavior to improve learning – with learning scaffolds and an evaluation of three key STEM competencies: mathematical competence (OECD, 2018), representational competence (Küchemann et al., 2021) – the ability to use various forms of representations to understand and communicate scientific concepts – and data literacy (Schüller et al., 2019; Kippers et al., 2018). Grounded in these educational research foundations, the system is designed to facilitate STEM learning by providing students with the opportunity to solve rapid assessment tasks (RATs). RATsApp employs both data-driven and expert-driven feedback to improve students' conceptual understanding. In addition to enhancing conceptual understanding, the system also aims to help students improve the aforementioned key STEM competencies. Additionally, the system provides a dashboard to students that displays their overall performance and their progress in the STEM competence levels. A lecturer dashboard is also available, allowing lecturers to monitor their lectures by tracking the progress of their students. Furthermore, RATsApp was developed in accordance with existing ethical guidelines for the development of digital technologies and systems (European Commission, Directorate-General for Communications Networks, Content and Technology, 2019; Abrassart et al., 2018; Systems & Software Engineering Standards Committee & Olszewska, 2021).

Moreover, RATsApp is designed to be easily expandable with machine-learning algorithms for automatic question selection, adaptive feedback, and provision of helpful materials. Currently, for security reasons, RATsApp's usage is exclusive to VPN users from two German universities: the Rhineland-Palatinate Technical University (RPTU) in Kaiserslautern-Landau and the Ludwig Maximilian University (LMU) in Munich. However, since the system is open-source, anyone can download the source code and implement it on their own server.

In this manuscript, we present findings on user experience derived from a usability survey based on the technology acceptance model 2 (TAM2) proposed by Venkatesh and Davis (2000). Due to subject availability, the user experience evaluation was conducted with a focus on higher education, hough the flexible design of RATsApp permits adaptation to other educational levels. The structure of the manuscript is as follows: we will first provide a comprehensive overview of the theoretical background, consisting of formative assessments, formative feedback, scaffolding, STEM competencies, learning dashboards, and our ethical considerations, detailed in Section 2. Following this, we explore the methods used to evaluate RATsApp in Section 3 and describe its capabilities in Section 4. In Section 5, we discuss the initial results of our evalua-

tion. Finally, we explore in depth our findings and their implications in Section 6. We then conclude our paper in Section 7 with a comprehensive summary of our research, limitations and possible lines of future work.

2 Theoretical background

2.1 Formative assessments

Formative assessments are used in the general classroom setting to diagnose students' difficulties and level of understanding (Shute, 2008; Bennett, 2011). These assessments are not intended to grade or rank students, but rather to provide feedback that can help them improve their understanding and skills. Such evaluations are important for educators to adapt their teaching methods and for students to improve their knowledge through formative feedback (see Section 2.2).

2.2 Formative feedback

Formative feedback is feedback that occurs during a learning process (Narciss, 2008). In general, formative feedback is known to play an important role in improving learners' understanding and enhancing their skills (Butler & Winne, 1995; Brookhart, 2017). Additionally, formative feedback activates learners' metacognitive processes, such as self-improvement and self-monitoring (Butler & Winne, 1995; Panadero, 2017). Hence, it is important to provide computer-based learning environments that are capable of providing formative feedback.

For formative feedback to be most effective, its content should answer the following three questions (Hattie & Timperley, 2007; Wisniewski et al., 2020): (1) Where am I going? (What content do I need to know?), (2) How should I proceed? (How much of the content do I know in relation to what I need to know?) and, (3) Where do I go now? (Where do I go next for more information?). Most importantly, Hattie and Timperley (2007), and later Wisniewski et al. (2020), argue that there are four different levels at which effective feedback can be provided: the task level, the process level, the self-regulatory level, and the self level. At the task level, feedback includes information about task performance, such as correct or incorrect responses. At the process level, feedback aims to address gaps in understanding of strategies for completing the task and identifying errors. The self-regulatory level involves feedback that enables learners to monitor and regulate their own learning. Finally, self level feedback emphasizes the learner's individual abilities and personal characteristics rather than focusing solely on the task, for example, a statement such as "You excel in mathematics". However, feedback on the latter level has been shown to have little or negative impact on students (Wilkinson, 1980; Kluger & DeNisi, 1998; Meyer et al., 1979), so it will not be taken into account here.

Additionally, it is important to note that the effective utilization of feedback, and thus RATsApp itself, is contingent upon the student's ability to accurately interpret and apply the information provided. Feedback, regardless of its appropriateness or

comprehensiveness, can only have a meaningful impact if it is understood and acted upon by the recipient (Ryan et al., 2022; Hattie & Jaeger, 1998).

One type of formative task-level feedback we use is known as *elaborate feedback* (Kulhavy & Stock, 1989; Shute, 2008). This type of feedback not only indicates whether an answer is correct or incorrect but also provides more detailed information on why the answer was correct or incorrect. This feedback is provided by experts in a specific and elaborated manner (Kulhavy & Stock, 1989; Shute, 2008). It contains an *evaluative component* (correct/incorrect information) along with an *informative component* that includes information related to the topic and specific details about the task. For example, it explains why the given answer was incorrect and provides guidance for students to arrive at the correct solution. Elaborate feedback has already been reported to be more effective than simply stating whether the answer is correct or incorrect (Bangert-Drowns et al., 1991; Pridemore & Klein, 1995).

The formative feedback mentioned above can be generated through two primary methods: expert-driven and data-driven. *Expert-driven feedback* is provided by individuals with a high level of knowledge or skill in a particular area (Deeva et al., 2021). *Data-driven feedback*, in contrast, is generated through the analysis of data (Deeva et al., 2021. The data analyzed in this way can be presented in a user-friendly manner, for example, through the use of dashboards.

2.3 Scaffolding

Scaffolding is an instructional technique that involves providing learners with support and tools to facilitate engagement with a task or problem (Puntambekar, 2022). These tools, known as scaffolds, can take various forms, including hints (Pea, 2004), prompts (Bannert & Mengelkamp, 2013), and explanations (Quintana et al., 2004). By offering an appropriate level of support at the right time, scaffolding can facilitate the gradual development of learners' understanding and skills (Hartman, 2001; Lin et al., 2012).

In education, scaffolding is used to fill cognitive gaps, which are the differences between a student's current knowledge and the domain knowledge expected at a certain point during their education. For example, if a student struggles to understand a course text due to his or her reading level, the teacher may use a didactic scaffolding approach, such as guided reading sessions, where teachers provide targeted support to help students understand texts at their current reading level. This approach gradually improves the student's reading skills until they can independently understand the assigned text without help (Van de Pol et al., 2015).

In summary, through explanation and guidance, students can develop the ability to effectively evaluate the accuracy of their responses (Wu & Puntambekar, 2012). In this way, scaffolding not only supports students in accomplishing tasks that might be beyond their individual capabilities, but it also cultivates new abilities and insights that empower them to handle similar tasks on their own in the future (Van de Pol et al., 2015).

2.4 STEM competencies

Three essential competencies in STEM learning are mathematical literacy, data literacy, and representational competence. One of RATsApp's objectives is to facilitate the development of these competencies in STEM education. These competencies are crucial not only for higher education but also for other educational levels. Additionally, they are important for the general public as they are prevalent in everyday life, from interpreting online information to making informed decisions in various fields. Below, we will elaborate on each of these competencies:

1. *Mathematical literacy* involves problem-solving skills, such as using mathematics to solve real-world problems by translating them into a mathematical expression, interpreting results, making predictions, and making reasonable judgments based on these results (OECD, 2018). For instance, a positive correlation between physics students' mathematical skills and their exam grades is well documented (Blumenthal, 1961; Hudson & Liberman, 1982; Korpershoek et al., 2015). Additionally, mathematical ability shows the strongest positive correlation with learning gain compared to prior knowledge and spatial ability in physics courses (Hake, 2002). This suggests that students with poor mathematical understanding may struggle in future physics classes. Similar results have been obtained for other non-STEM disciplines, such as economics (Ballard & Johnson, 2004; Llamas et al., 2012) and plant psychology (Pozo & Stull, 2006).
2. *Data literacy* is defined as the combination of efficient behaviors and attitudes that enable individuals to effectively carry out all the steps involved in creating value or making decisions from data (Schüller et al., 2019). It is a foundation for data-based decision making (Kippers et al., 2018) and a crucial competence in the 21st century, as it enables a person to analyze and interpret data, to provide others with relevant data, and it facilitates the systematic generation of knowledge based on data (Schüller et al., 2019). Data can be presented in various forms, such as tables or graphs. The understanding of graphs is a prerequisite for learning in most subjects in higher education (Bowen & Roth, 1998).
3. *Representational competence* involves the ability to extract information from representations, such as graphs, tables, diagrams or images , to translate between different types of representations, and to interpret and construct representations (Küchemann et al., 2021). Research has shown that students' problem-solving performance is strongly influenced by their representational competence (Kohl & Finkelstein, 2006; Edelsbrunner et al., 2023).

A first step to increasing the three STEM competences mentioned above is to provide learners with feedback on their current level of competence based on their performance.

2.5 Learning dashboards

Learning dashboards serve as a tool to facilitate learners' comprehension of their progress and performance through the provision of easily interpretable visualizations of their data (Verbert et al., 2013, 2014; Duval et al., 2012). For instance, if a student

wants to check their progress for the week, they could select a “Progress" tab on the dashboard. This action would reveal a chart that illustrates their accomplishments for each day of that week. These dashboards capture and display traces of learning activities with the aim of promoting awareness, reflection, and sense-making (Duval et al., 2012).

Additionally, they enable learners to establish goals and monitor their progress towards achieving them (Verbert et al., 2014). A student could, for example, establish a target to enhance their rate of learning in the upcoming month, and utilize the dashboard to monitor their daily advancement towards this target. Research has demonstrated the utility of dashboards in different educational contexts. For instance, Toyokawa et al. (2023) found that dashboards for high school students were effective in facilitating foreign language reading. In addition, the use of dashboards has been shown to help students manage their time effectively and implement new strategies to facilitate learning in higher education (Chen et al., 2023).

Additionally, dashboards may contain data on students' performances to support teachers in diagnosing student difficulties (Knoop-van Campen et al., 2021). They allow teachers to rapidly identify areas in which students may require additional support and to adjust their pedagogical approach or select an appropriate learning activity accordingly Van Leeuwen et al. (2014). For instance, after noticing that a student consistently performs poorly on quizzes, a teacher might choose to incorporate more subject-specific activities into the teaching plan.

2.6 Technology acceptance model 2 (TAM2)

To understand technology acceptance and adoption, several models and surveys provide complementary perspectives and insights to understand the factors that influence technology use. One such model is the technology acceptance model 2 (TAM2) proposed by Venkatesh and Davis (2000). This model provides a framework for predicting user acceptance of technology. TAM2 integrates two main categories: social influence processes and cognitive instrumental processes.

Social influence processes in TAM2 include subjective norms, voluntariness, and image. Subjective norms refer to an individual's perception of whether significant others believe they should use the new system. Voluntariness refers to how voluntary the use of the technology is perceived to be. The image reflects how using the technology enhances one's status or image within a social group.

Cognitive instrumental processes include job relevance, quality of output, demonstrability of results, and perceived ease of use. Job relevance refers to the extent to which the technology matches the user's job tasks. Quality of output refers to the extent to which the technology produces outputs that meet quality standards. Demonstrability of results refers to the extent to which the results of using the technology are observable and tangible. Perceived ease of use is the extent to which an individual believes that using the technology will be effortless, influenced by personal experience and the intrinsic characteristics of the technology.

TAM2 proposes that perceived usefulness and perceived ease of use are the primary factors influencing a user's intention to adopt a technology. To understand these

influences comprehensively, it is necessary to analyse all the constructs of TAM2 and their relationships with perceived usefulness, perceived ease of use, and ultimately, the intention to use the technology. Technology developers can benefit from understanding TAM2 when designing systems that are more likely to be accepted by users. By addressing these factors, they can increase user acceptance.

2.7 Ethical considerations

It is important to consider ethical aspects during the development of educational feedback systems. However, at the moment, this is still rare. In this section, we briefly introduce the ethical guidance for RATsApp.

Prior to the development, we reviewed existing ethical guidelines for the development of digital technologies and systems (European Commission, Directorate-General for Communications Networks, Content and Technology, 2019; Abrassart et al., 2018; Systems & Software Engineering Standards Committee & Olszewska, 2021). Some of them focus specifically on the teaching and learning context (European Commission, Directorate-General for Education, Youth, Sport and Culture, 2022). We are critical of these guidelines insofar as they are often too general or schematic or do not refer to the teaching and learning context (with the exception mentioned above). Furthermore, they all refer primarily to technical aspects, such as robustness, or concentrate on questions concerning data security. These belong rather to the legal than to the ethical sphere. Hence, the underlying image or concept of man, in the philosophical sense, within the practical field of learning and teaching is also disregarded (Kemp, 1992; Zichy & Alber, 2017; Joisten, 2020b). Given this background, the ethical orientation - that enables the ethical development of a product in the first place - is currently missing.

Our ethical approach aims to reflect the practical field of learning and teaching at university in STEM subjects - which are in the introductory phase - with the help of *Technoethics for Emerging Digital Systems* (TEDS) (Joisten et al., 2022). This approach considers the various ethical challenges that may arise in relation to digital technologies, and incorporates different ethical modes of access. To put it more precisely, the implementation of a new technology within a specific context, such as a learning environment, can only be effectively guided by diverse perspectives of a complex technological ethics, as exemplified by TEDS. Areas of TEDS enable this ethically appropriate perspective concerning: *Anthropo-*, *Pre-*, *Meta-*, *Principle-*, and *Applied*-Technoethics.

These approaches can be summarized as follows: the view of humanity or the concept of man, that guides the field of practice, is fundamental to Anthropo-Technoethics. If we pay attention to the practical field of learning and teaching, it can be noted that humans have the capability to learn in order to become (and this is a normative assumption) more autonomous and to improve their ability to think critically. If this capability is developed with the help of lecturers and peers, it becomes an inherent capability that can also be integrated into a cultural and social context under favorable external circumstances (Nussbaum, 2009; Sen, 1997, 1987). From the perspective of Anthropo-Technoethics, this implies that the extent of the impact of the technical

product on learners must be determined and analyzed in terms of its impact on the human being. The guiding question is whether or not the system facilitates or impedes individual autonomy, as well as the user's motivation and interest. However, successful learning is often perceived as the mere acquisition of knowledge. As a result, the effectiveness of digital learning systems is mostly assessed in terms of this knowledge acquisition without considering other abilities and competencies that are also important to associate with it. Anthropo-Technoethics addresses this gap.

The pre-technoethical perspective focuses on the technical product, in our case the feedback system, which is based on short tests. This approach aims to uncover any potential ethical concerns while the development of the system itself is still in progress, as digital technologies may already harbor ethical problems. The issue of inadequate reflection is evident, for instance, in computer programs such as COMPAS[1], which predicts the likelihood of recidivism or risk rates for offenders in the United States and exhibits discriminatory practices towards people of color. When judges base their judgements on the program's calculated predictions, they transfer normative decisions to an algorithm that is proprietary and lacks transparency. This software and the output of the used algorithm contains proven racial and age-related biases, which should have been identified during development, or the system should have been discontinued as soon as these biases were recognized (Engel et al., 2024). By identifying and analyzing such ethical issues within an interdisciplinary team, digital technologies can be technically adapted and enhanced. Considering the current state of the project, such technical issues are not identified. It is a transparent, open-source system that students use voluntarily, and it does not make significant decisions that could profoundly affect their lives, such as denying access to exams or lectures.

Meta-Technoethics pursues theoretical considerations at a more abstract level. For example, inappropriate connotations can be exposed and problematized here by analyzing the use of terms for digital technology. The objective of this perspective is not only to deconstruct (self-evident) concept transfers of human qualities to digital systems, but also to find more appropriate terms. An example, in the context of AI, is the expression 'decision-making'. This is an anthropomorphism, since digital systems do not make decisions in the ethical, philosophical sense of the word, but present calculated probabilities (Kaminski, 2020). In the current state of the project, category errors are not given in the meta-technical sense.

Applied-Technoethics is mostly concerned with technology assessment. One of its main tasks is to analyze the potential and effects of scientific as well as technological developments comprehensively and with foresight, and to explore the associated social, economic, ecological opportunities and risks, as described in the archive of the German Bundestag[2]. In our project, we focus on technology assessment. It is the central task of Applied-Technoethics to critically examine the respective digital application within a concrete context. For example, it is conceivable that the impact on students working with RATsApp reveals unforeseen anomalies that have to lead to a correction of the application. This implies that Applied-Technoethics principally

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] https://www.bundestag.de/webarchiv/Ausschuesse/ausschuesse19/a18_bildung/technikfolgenabschaetzung

centers on the assessment of technology to ultimately facilitate modifications to the application.

In our project Principle-Technoethics is an important foundation. Principles are the normative basic orientations in the learning context. They can be defined as basic values, basic beliefs, and targets (to be realized) that guide the action in a field of practice (Beauchamp & Childress, 2019). In the teaching and learning context, the following *primary principles* related to learners can be identified in a topical method, among others (Joisten, 2004, 2006): development of autonomy, problem-solving behavior, independence, voluntariness, (self-)responsibility, learning success, learning how to learn, cooperative learning (ernately helping and receiving help), and creativity. These must be taken into account for optimal learning in the interaction between students, lecturers, and RATsApp. If the development and application of RATsApp are conducted in accordance with the aforementioned principles, the extent of learners' alignment becomes evident. This, in turn, eventually leads to an attitude that enables them to increase their critical self-reflection (Joisten, 2020a). In the sense of a Principle-Technoethics approach, the best possible education for humans and thus ethical action - with the assistance of new technologies - can succeed only with an appropriate consideration of these principles. We aim to take a first step in this direction with RATsApp.

2.8 Research aims

Given the pressing need for cost-effective educational learning systems that can also support large-scale research, we have developed an open-source STEM AFS. This system is freely available for download and modification, and it can be utilized across various STEM educational levels. It is based on the key research-based features mentioned earlier. In this article, our primary objective is to assess student perceptions and usage of RATsApp within the context of university STEM education. hough RATsApp could be implemented at various educational levels, we are focusing on higher education due to our access to students in this setting.

In order to assess the usability of RATsApp, we use a theoretical extension of the technology acceptance model developed by Venkatesh and Davis (2000). This approach will enable us to analyze and interpret survey data from students, thereby providing insights into crucial factors such as the system's perceived usefulness, its ease of use, and students' behavioral intentions towards using it. This type of usability study can be a powerful tool for the identification of areas for improvement and the enhancement of the overall user experience and satisfaction.

3 Materials and methods

3.1 Participants

The system was implemented as a stand-alone version in the winter semester 2022/23 at two German universities: the RPTU Kaiserslautern-Landau and the LMU Munich.

At the RPTU, it was tested in four undergraduate-level lectures: Higher Mathematics for Civil Engineers I, Fundamentals of Mathematics II for prospective mathematics teachers, Fundamentals and Applications of Probability Theory and Mathematics for Economists. At the LMU, RATsApp was tested in three first-semester undergraduate lectures: Experimental Physics 1, Experimental Physics for Veterinary Students and Mathematical Methods in Theoretical Physics. These lectures were conducted in person. At the beginning of the semester, a representative from our team presented RATsApp and its functionality in the lectures. All the RATs were used to support normal lecture assignments and were not considered in the final grades for the courses. The students accessed the website outside of lecture hours and during their free time. At the end of the semester, we revisited the lecture halls of the aforementioned courses and requested voluntary participation to evaluate RATsApp. The survey was conducted during these visits. A total of 64 people who attended these lectures participated in a survey to evaluate usability, as described in Section 3.2.

To maintain participation throughout the semester with the final purpose of testing user experience at the end, we provided monetary incentives to active users. Specifically, we held five lotteries during the semester, each offering 15 prizes of 20 Euros per lecture. The criteria for lottery participation varied each time: for the first lottery, students needed to complete a high school level STEM test; for the second, completion of RATs, which are rapid assessment tasks tailored to each lecture, during the first four weeks was required; for the third, completion of RATs during weeks five to ten was necessary; for the fourth, completion of RATs from week 11 onwards was required; and for the fifth, students had to either report their final exam grade or provide a reason for opting out. The lotteries were based solely on participation and did not require fast or correct responses. The validity of the high school level STEM test and the RATs is not within the scope of this manuscript; they are mentioned solely to clarify the participation criteria.

3.2 Usability evaluation

To measure students' user experience with the system, we created a usability questionnaire based on the TAM2 model (Venkatesh & Davis, 2000) (see Section 2.6). Our survey is focused on the cognitive instrumental processes section of the TAM2, particularly the constructs of "Intention of use", "Perceived Ease of Use", "Perceived Usefulness", "Job Relevance" and "Output Quality". We translated these sections to German and adapted them to the context of RATsApp. For example, we renamed the "Job Relevance" section to "Relevance to the Course of Study". In addition to the TAM2 sections described above, we included questions about age, gender, frequency of use, attended lecture, and whether participants had a RATsApp account, as having an account was not a requirement for completing the survey. The distinction regarding account status is clarified in the results Section 5. The questionnaire we used can be found in the Appendix A. As mentioned in Section 3.1, the survey was voluntary and conducted in the lecture halls. It was administered online, allowing students to participate using their mobile phones by scanning a QR code or using a provided link. ernatively, students had the option to complete the survey at a later time if they

preferred. The survey was anonymous and did not require participants to use their RATsApp accounts. More details about the lectures that took part in the survey can be found in Section 3.1.

4 Functions and features of RATsApp

The system has been designed to accommodate four distinct user types: student, RAT creator, lecturer, and administrator. We differentiate between RAT creators and lecturers, as the individuals creating the educational content may differ from those delivering it. For example, in our implementation, undergraduate teaching assistants created the RATs, which were then evaluated by the lecturer. The specific capabilities of each user type will be elaborated upon in Section 4.4. In accordance with the architectural design proposed by Liu and Yu (2022), the system is composed of three layers, as shown in Fig. 1: the presentation layer, which handles the user interface; the system layer, which manages the website's logic and functionality; and the database layer, which stores and retrieves data.

4.1 The presentation layer

The presentation layer of the website is responsible for rendering the user interface. To enhance the user experience, front-end technologies such as HTML, CSS, Jinja2, and JavaScript have been employed. The website has been designed to exhibit responsiveness and scalability across various devices, including computers, tablets, and mobile

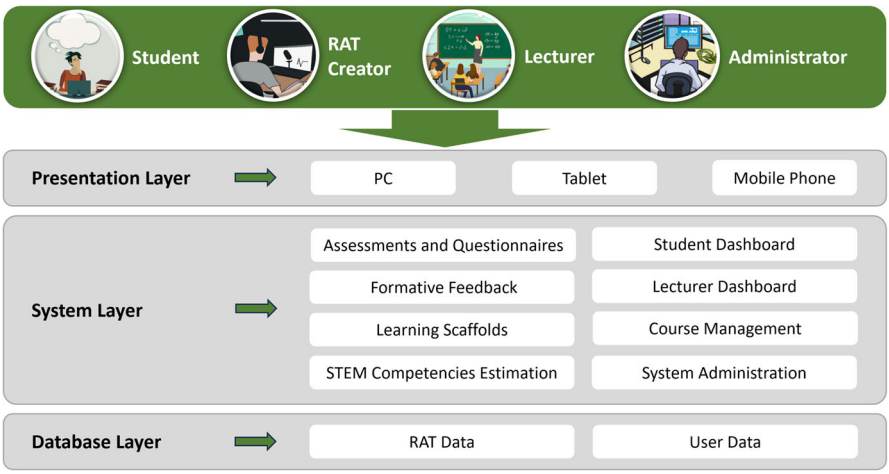


Fig. 1 RATsApp supports four distinct user types: student, RAT creator, lecturer, and administrator. The architecture of RATsApp is divided into three layers: the presentation layer, the system layer, and the database layer. Each layer serves a specific function in the overall operation of the website

phones. Additionally, the interface style was tested by our developers across the most commonly used web browsers to ensure a consistent and positive user experience.

4.2 The system layer

4.2.1 RATsApp assessments and questionnaires

In this system we test the abilities of students by using assessments called RATs, which are short and quick formative assessments (see Section 2.1). In Fig. 2, we present an example of how students view a RAT. The RAT shown in the figure is from an undergraduate course in Experimental Physics 1, which we have integrated into our

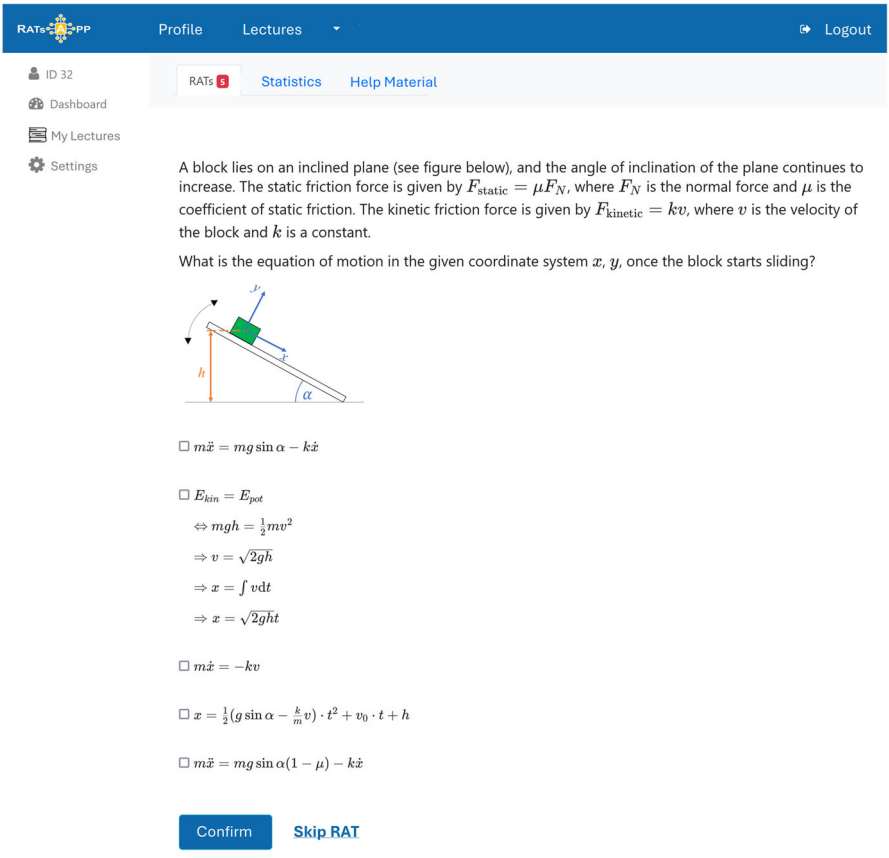


Fig. 2 Snapshot of a student's view of the RATsApp website while answering a RAT. The page features a top navigation bar with access to the student's profile, available lectures ("Lectures") accessible with a registration password, and a logout option. The sidebar provides access to the student dashboard, registered lecture materials ("My lectures"), and settings. There are three additional tabs: the "RATs" tab, which displays the current RAT and shows a red rectangle indicating the number of unanswered RATs remaining for the current sheet; the "Statistics" tab, which provides RAT statistics; and the tab "Help Material" for scaffold support related to the RAT they are answering

system. In our implementation, RATs were created using the recommended materials of the lectures they are intended for, as well as other sources such as books and the internet. The creator of the RAT assigns competence criteria, scaffolds, topics, and concepts to it. Before it is made available to students, each RAT must be approved by two users with elevated user rights to ensure its correctness. In our implementation, for example, RATs were created by undergraduate teaching assistants and were approved by two subject experts, specifically PhD students specializing in the relevant topic or the lecturer.

RATs can be utilized to generate exercise sheets, questionnaires, tests, and cross-lecture questionnaires. RATs that belong to the cross-lecture questionnaires can be used to test knowledge all students benefit from, regardless of their course of study or attended lectures. For instance, these RATs may be used to evaluate various high school proficiency levels of first-year university students, serving as a diagnostic tool to assess their initial knowledge base.

4.2.2 RATsApp feedback

RATsApp aims to provide formative feedback that is useful for students to narrow the gaps between their understanding and the curriculum. To enhance the learning experience, the system provides constructive feedback as students complete each RAT. This feedback has to be developed by experts in the field, such as the lecturers, and structured using elaborated feedback techniques that incorporate both evaluative and informative components, as outlined in Section 2.2. This feedback is provided immediately following the completion of a task, providing students with explanations as to why each option is correct or incorrect (see Fig. 3). The informative component of this feedback could include the following information: i) an explanation to guide the student in identifying why the chosen answer is incorrect; ii) information to guide the student through steps on how to solve the task correctly; iii) information to guide the student to external scaffolds; or iv) information that includes a combination of all or some of the above options. In our implementation, the feedback was created by PhD students with expertise in the field or the lecturers themselves. The feedback creator determines the most relevant information to provide and the type of feedback needed based on the task. Therefore, this feedback helps students at the task level and/or the process level, depending on the difficulty of the task.

4.2.3 RATsApp scaffolds

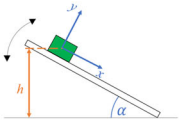
During the task, students are offered optional scaffolding support (refer to Section 2.3). This support is not automatically provided through pop-ups or similar means; rather, students must actively access it by clicking on a tab adjacent to the task (see Fig. 4). The scaffolds are related to the topic and concept of the task. It can be provided in the form of a text, a link to external online material, or a reference to learning material. We argue that this helps students who are at a self-regulatory level, i.e., they are able to control their learning process (see Section 2.2). Unfortunately, students who are not at a self-regulatory level may have difficulty benefiting from the optional scaffolding support because the scaffolding requires active student selection.

Correct!

The question:

A block lies on an inclined plane (see figure below), and the angle of inclination of the plane continues to increase. The static friction force is given by $F_{\text{static}} = \mu F_N$, where F_N is the normal force and μ is the coefficient of static friction. The kinetic friction force is given by $F_{\text{kinetic}} = kv$, where v is the velocity of the block and k is a constant.

What is the equation of motion in the given coordinate system x, y, once the block starts sliding?



Evaluation and Feedback on the Options:

The table below gives you feedback on your answers.

- The answers for which you have ticked the box, i.e. you have agreed with the statements, are marked in **orange**.
- The **first column** indicates which of the option statements are correct or incorrect.
- The **third column** gives you feedback as to why the option statement in **column 2** is correct or incorrect.
- In **the last column**, you have the opportunity to leave comments for us on each option.

Option Correctness:	Option Statement:	Feedback on the Option:	Comments on the Option:
✗	$m\ddot{x} = mg\sin\alpha(1-\mu) - k\dot{x}$	This is incorrect because the static friction force $F_{\text{static}} = \mu F_N = \mu mg\cos\alpha$ ceases to act once sliding begins.	Enter your comment here
✗	$x = \frac{1}{2}(g\sin\alpha - \frac{k}{m}v)\cdot t^2 + v_0 \cdot t + h$	This is incorrect. It does not represent an equation of motion, as it fails to account for acceleration. Moreover, it is not a complete solution to an equation of motion since $v = \dot{x}$ remains on the right-hand side. Additionally, the initial position in the coordinate system should be zero, not h.	Enter your comment here
✗	$E_{kin} = E_{pot}$ $\Leftrightarrow mgh = \frac{1}{2}mv^2$ $\Rightarrow v = \sqrt{2gh}$ $\Rightarrow x = \int v dt$ $\Rightarrow x = \sqrt{2gh}t$	This is incorrect because the energy conservation equation $E_{\text{kin}} = E_{\text{pot}}$ does not hold in the presence of friction.	Enter your comment here
✗	$m\dot{x} = -kv$	This is incorrect because the equation fails to specify the resulting force and erroneously implies that $\dot{x} = \frac{k}{m}v$.	Enter your comment here
✓	$m\ddot{x} = mg\sin\alpha - k\dot{x}$	That is correct. The resulting force is calculated from the downhill force $mg\sin\alpha$ and the opposing kinetic friction force $F_{kinetic} = k\dot{x}$. The static friction force does not need to be considered here since the body is already in motion.	Enter your comment here

Fig. 3 Snapshot of the feedback provided to the student after solving the RAT. The top part of the page indicates whether the RAT was solved correctly or incorrectly, with correct solutions highlighted in green and incorrect ones in red. Below, the text of the RAT is displayed again, followed by a table providing detailed feedback. Instructions on how to interpret the table are given above it. The table includes: an orange rectangle highlighting the student's selected answers; the first column indicating whether each option statement is correct or incorrect with a green check mark for correct answers and a red cross for incorrect ones; the second column containing the text of each option; the third column providing feedback on why the option is correct or incorrect; and the last column allowing the student to give feedback on each option

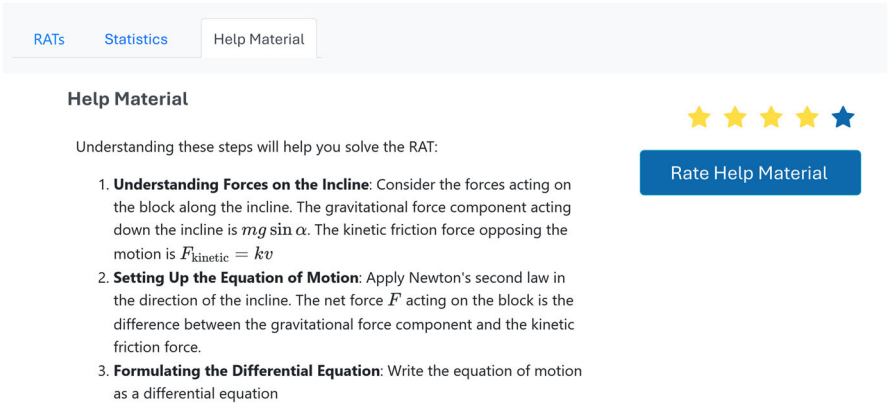


Fig. 4 Snapshot of a text scaffold example from the "Help material" tab, from Fig. 2. The scaffold support can be provided as text, a link to external online resources, or a reference to learning materials. Students also have the option to rate the scaffold support

From an ethical perspective, it is important to emphasize that hough an active decision must be made to access the provided learning materials, it is still essential to consider their ease of access. As a result, students do not experience irritation, frustration, or, in Wiegerling's words, resistance (Wiegerling, 2021). Such a resistance can be highly beneficial in the learning process if learners can overcome it themselves. For instance, when students realize they do not understand a formula or cannot solve a problem (which constitutes experienced resistance), they may seek help from another student or a teacher, consult relevant books for solutions, and so on. By trying to resolve such challenges independently, students can attribute the credit for their individual learning progress to themselves. This effort to overcome resistance may cultivate an attitude in the ethical sense, enabling individuals to attribute the credit of self-learning to themselves, even in the face of difficulties. This does not only enhance their problem-solving skills but also deepens their self-confidence (Fuchs, 2020).

4.2.4 Skill level estimation of STEM competencies

We aim to estimate three key STEM competencies: mathematical literacy, data literacy, and representational competence, as defined in Section 2.4. To achieve this, we developed a list of criteria, with each criterion linked to one or more of these competencies. We categorized each RAT using this list to determine which STEM competencies are required to solve it. The representational competency part of the list was created based on the criteria described by Scheid and colleagues (Scheid et al., 2017). Furthermore, we used the definition and criteria of the mathematical literacy of the Programme for International Student Assessment (PISA) 2022 mathematics framework (OECD, 2018) and the definition and criteria of data literacy (Schüller et al., 2019). In order to make a list suitable for evaluating a large number of RATs, we have compressed the information into 21 items (see Appendix B). Please note that the list of questions needs to be validated in further studies, as this is beyond the scope of this manuscript.

Upon completion of a RAT by a student, the system estimates their competence levels based on the required level of each competence needed to solve that RAT. If the RAT was solved correctly, the score for each competence is added to the student's current score for that competence. The student's updated scores for each competence are then normalized by dividing them by the maximum possible score, calculated from the total potential points of all completed RATs by that student. A graph visualizes the connection between competence level and number of solved RATs for each competence after every completed RAT. This feedback is intended to help to identify weaknesses and strengths in one or more of these competencies. This is for the purpose of guiding the students on what kind of tasks they should work on more to ultimately improve their overall academic performance. In the future, a machine learning (ML) algorithm could be trained to automatically assign those RATs to students that could help them improve a specific competence.

4.2.5 Student dashboard

A user-friendly dashboard displaying key performance metrics is also available to students for tracking their progress. The task performance statistics provided in the dashboard include the percentage of tasks answered correctly or incorrectly and the total number of RATs answered per week. This data-driven feedback aims to help students to set specific learning goals and monitor their progress over time. This process helps to activate the metacognitive processes necessary for self-regulation, such as planning, monitoring and evaluating.

From an ethical and didactic standpoint, it is crucial to consider the potential for demotivation among students when confronted with transparent data regarding their individual task completion statistics (Hattie, 2008). As already mentioned (see Section 2.2), a certain level of competence in interpreting and handling feedback is required. To address this, one can implement training sessions and provide guidance to help students develop the necessary skills to effectively interpret their data.

4.2.6 Lecturer dashboard

Lecturers have access to dashboards that display students aggregated statistics and difficulties. They can monitor their students' performance on RATs and individual competencies. This feature allows lecturers to conduct formative assessments and possibly evaluate the effectiveness of their teaching methods by identifying patterns in student performance. When analyzing the data, lecturers can compare the performance of different student groups, track progress over time, and correlate specific teaching methods with improvements or persistent challenges. The dashboard also displays RATs that are frequently answered incorrectly by students, which are displayed in the following error categories:

- Always answered incorrectly.
- Often answered incorrectly (answered correctly less then 40% of the time).
- Deceptive answer (the same wrong answer was selected more then 30% of the time the question was answered).

In addition, lecturers can see the most frequently chosen answer option. This can help them identify misconceptions their students may have and intervene. This feedback can also contribute to the improvement of the teaching methods in a self-reflective way, selecting new approaches, and ultimately implementing them.

4.2.7 System administration

A system administrator is responsible for granting access rights to new users. Upon request, an admin can also delete an account while retaining non-personal data and log entries. Furthermore, system administrators also receive data-driven feedback from the system. They have access to statistics for all available lectures. In addition, system administrators can see how many RATs are generated per lecture per week and which user ID created them. Also, they can see how the users interact with the site by monitoring log entries.

4.2.8 Course management

This component of the system layer is responsible for managing all course-related content. This includes the creation of RATs, the generation of weekly RAT sheets, and the development of lecture materials, among other tasks. These features are further explained in Section 4. The system is designed to facilitate the efficient organization and delivery of course content to enhance the learning experience for students.

4.3 Database layer

The database architecture of RATsApp is divided into two distinct components: the RAT database and the user database. The RAT database stores the data required for the website's operation, while personal user data is stored separately in the user database to mitigate the risk of inadvertent exposure due to programming errors. Security is a paramount concern for the system, and as such, widely used and well-documented frameworks are employed for server-side code development and database access. RATsApp is written as a Python web application using flask[3] and SQLAlchemy[4]. It runs on Gunicorn[5] as a wsgi server and is served from the web using nginx[6] as a reverse proxy.

These frameworks are actively maintained and provide inherent protection against common security vulnerabilities, including SQL injection attacks (Lerner, 2013). MariaDB[7] is used as an SQL database to save the data. Furthermore, the server is only accessible from the networks of the universities of LMU Munich and RPTU Kaiserslautern-Landau as well as from the corresponding VPN connections. This is meant to lower the number of potential intrusions from outside.

[3] https://flask.palletsprojects.com

[4] https://www.sqlalchemy.org/

[5] https://gunicorn.org/

[6] https://nginx.org/

[7] https://mariadb.org/

4.4 Users

When registering, RATsApp only requires a valid email (belonging to one of the institutions from which registration is accepted) and a password. In addition, the user must accept the terms of use and choose one of the following four user categories: students, RAT creators, lecturers or administrators, with user access rights increasing in this order. We differentiated between RAT creators and lecturers to provide an option for lecturers to have the RATs for their lectures developed by individuals such as graduate teaching assistants. Below we will explain the different categories of users and what content they can access. We start with users with a low number of access rights (Students) and continue in ascending order. To differentiate between different user types and keep track of the interaction history, we created a comprehensive account system that allows the users to access features commonly found on websites, such as email validation, signup, or changing the password.

4.4.1 Students

Students can register for a lecture by first searching for the lecture in the systems course catalog and then entering the lectures password, which is usually provided by the instructor. In their profile, students have access to a sidebar in which they can select the lecture to work on (see Figs. 2 and 5), or they can access a cross-lecture questionnaire available to them. A cross-lecture questionnaire is an option only available to students and not for other user types. It can be assigned to them, for instance, to test fundamental knowledge that all students benefit from, independent of the lectures they attend.

When a lecture is selected, students can view the RATs that were either automatically selected or selected by the lecturer manually, RAT sheets or live RATs. The latter are only accessible when a live session is in progress. More details about these options can be found in Section 4.4.3. The topics of the RATs can also be limited by a topic selection function. A student accessing any of the above options, will be presented with the first available RAT, as shown in Fig. 2. In addition, at the top of the RAT a tab bar will appear that allows the student to navigate to the overall statistics of the answered RATs or to display and, optionally, rate scaffolds (see Fig. 2). As explained in Section 4.2.5, the student statistics presented in the tab include the percentage of RATs answered correctly/incorrectly and the total number of RATs answered per week. In addition, as explained in Section 4.2.3, the scaffolds provided in this tab can be suggested feedback from experts in the form of text, a link to a video, a link to external online material, or a reference to a book. This suggested material is related to the topic and concept of the RAT that the students are working on.

After answering a RAT, the student receives elaborated feedback created by an expert (see Section 4.2.2 and Fig. 3). Furthermore, students can suggest scaffolds they found helpful in answering the RAT, leave comments, or point out possible errors in the RAT. Afterwards, the student moves on to the next RAT or, if the RATs were completed, is shown the newly updated competence levels presented in Section 4.2.4. This is a way to determine if a student is having difficulties with any of the skills

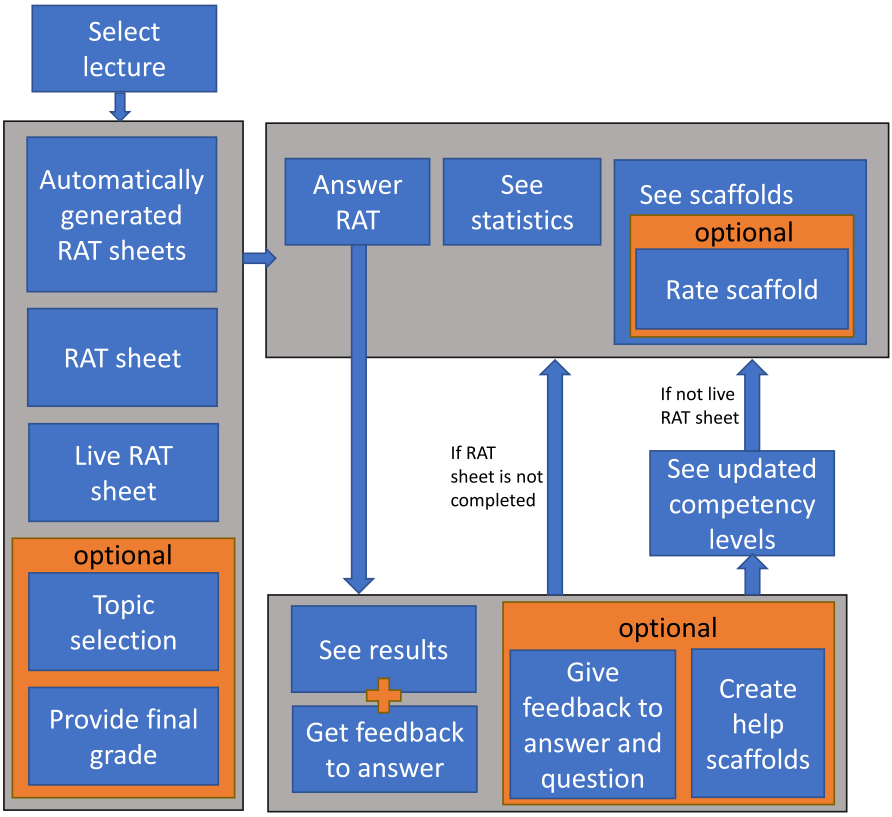


Fig. 5 *The student workflow of answering RATs:* first a lecture has to be selected, afterwards a student can choose from three options: answering an automatically generated RAT sheet, a RAT sheet, or a live RAT sheet. This will lead to a new page with the choice to answer the RAT, view scaffolds or statistics. After answering a RAT, the results and feedback are shown. The student is then either guided to the next RAT in the sheet or shown their updated competence levels if the sheet is finished

associated with specific competencies. The complete student workflow is visualized in Fig. 5.

4.4.2 RAT creators

A RAT creator can be a higher education teaching assistant, an administrator, or a lecturer. All RATs can contain the following information:

1. **Subject/Topic:** A subdivision of the lecture such as ”Matrix operations” in a Mathematics lecture. With this thematic category of the RAT, they can be searched in the database. Hence, the RAT can be used in different lectures. In addition, this category specifies the knowledge required to solve the RAT correctly.
2. **Concept:** A concept is a more specific label within the subject, such as, in the case of the subject “Matrix operations", a related concept could be “matrix mul-

tiplication". In other words, concepts are a more fine grained subdivision of a topic.

3. **Lecture:** At the moment of creation, the RAT can be assigned to one or more lectures which enables it to occur in automatically generated questionnaires if all connected topics and concepts have been taught in the lecture.
4. **Type of question:** The RAT creator can select three types of questions; multiple choice, multiple True/False, or open-ended.
5. **Answer specific or elaborated feedback:** When creating the right/wrong answers of the multiple choice, the RAT creator can add individual feedback to the specific answer. As explained in Sections 4.2.2 and 4.4.1, this is with the purpose of helping the students to understand why a certain option was the incorrect/correct one.
6. **Competence category:** The RAT creator can label it by choosing options from a pre-determined list (see Appendix B) that helps to estimate how much of a specific competence is needed to correctly solve the RAT. The considered competencies are mathematical literacy, data literacy and representational competence.
7. **Scaffolds (Optional):** As explained in Sections 4.2.3 and 4.4.1, scaffolds can be given to the students related to the subject/topic and concept of the RAT. These scaffolds can be in the form of text, links to videos, links to external online material, or book references. Students can view the scaffolds while solving a RAT. Unlike other items in this list, scaffolds are optional and can be added to a RAT at a later time.
8. The creator of the RAT may select an additional option, by assigning it to be part of a **cross-lecture questionnaire** (see Section 4.2.1). If a RAT is marked as belonging to a cross-lecture questionnaire it can not occur in regular RAT sheets or as an automatically selected RAT.

A powerful embedded text editor (Quill[8]) is used to create the RATs and the scaffolds. This editor supports latex and images. Once a RAT is created, it can be edited, duplicated, deleted, revised (see below), and searched by author, lecture, topic, or concept. Since the quality of a RAT is of great importance, it is not immediately available to students at the time of its creation. This is the case even when a RAT is linked to a lecture in the database. To maintain quality control of the RAT, a review function was implemented. In this option, at least two experts in the field, such as the lecturer or PhD students, must approve the item along with the corresponding feedback for it to be used in the lecture. The required number of experts can be adjusted in the configurations. The reviewer has the opportunity to communicate with the creator of the RAT through a comment function in order to clarify any doubts or possible errors. The RAT creator will receive a notification email when comments are made on the RAT. The email includes the RAT id and the comment itself. Anyone who makes a comment will receive an email notification of new comments. This reviewing process ensures that RATs used in the automated system (explained in the next Section) exhibit a good quality.

[8] https://quilljs.com/

The same applies to the scaffolds. Its content must be approved before it is made available to the students. However, it only has to be approved by one reviewer (this number can be changed in the configurations).

4.4.3 Lecturers

During the creation of the lecture, the following information should be entered: the name of the lecture, the intended audience, the scheduled dates and times, lecture term, and a lecture code or password for students to join. Additionally, a syllabus can be created using the the scheduled dates for the lecture by adding subjects and concepts to be covered on a specific date. This is necessary to automatically add questions to a test because the system needs to verify if the required concepts to solve a RAT have already been taught.

In order to assign RATs to the students (see Fig. 5) the lecturer has the following options:

1. The lecturers can use the syllabus to **automatically** generate a RAT sheet, which is a set of RATs related to the subjects and concepts included in the syllabus. The RAT sheets can be viewed in the student's profile starting from the assigned date. In this option, the lecturer also has the opportunity to see the pool of possible RATs and remove undesired RATs from the sheet.
2. An additional function for adding RATs **manually** is to select them from a list of RATs belonging to the specific lecture. The created RATs are displayed in the same place as those automatically selected by the system.
3. The lecturers can also create **RAT sheets**, which are collections of RATs, analogous to exercise sheets typically assigned as homework in a given lecture. These sheets are created with a unique name, and lecturers can select RATs from a pool of existing ones and arrange them in the desired order of appearance.
4. Lecturers also have the option to use the website for **live RAT sheets**. This lets them select an already created RAT sheet and open a live session where all students in the lecture are able to answer the RATs simultaneously. Statistics describing the student's performance for each individual RAT and the whole RAT sheet are shown in real time to the lecturers. This feature enables active learning, a teaching method that can increase interaction between students and lecturers, as well as between students themselves, and thereby provides formative feedback in real time (Freeman et al., 2014).

As mentioned in Section 4.2.6, lecturers also have access to the statistics of students in a lecturer dashboard, including metrics such as average student performance on RATs.

4.4.4 Administrators

An administrator has full access to all areas of the website, including the additional capabilities outlined in Section 4.2.7. These capabilities include managing user accounts, moderating content, and configuring site settings, among others.

5 Results: Usability survey evaluation

To evaluate the attitude of students towards the website and its output quality, we created a usability survey based on the TAM2 framework which was conducted at the end of the winter semester 2022/23 (see Section 3.2). Participation in the survey was voluntary and anonymous, as there was no need to log into RATsApp for it (see Section 3.1). The only personal information asked for in the survey was optional and consisted of age and gender. In order to further categorize the users, the test included a field to indicate a lecture they attended and whether or not they had created an account in RATsApp. We categorized students as non-users if they stated that they had never used the website even if they were registered. In total, eleven registered users stated that they never used the website and were therefore counted as non-users. Taking this into account, the survey was completed by 36 users and 28 non-users. The average age of the participants was 19.87 with a standard deviation of 3.19 years. Of the participants, 28 were male, 33 were female, and three provided no information.

The usability test consisted of three elements: multiple choice questions, a Likert-type questionnaire and written feedback (see Table 2 in Appendix A). One of the multiple-choice questions relevant to this survey was related to frequency of use, as shown in Fig. 6. The most frequent responses to the usage question were "Twice a month" and "Once per semester," each accounting for approximately 25% of total responses. Only around 5% of respondents reported daily usage. The median usage value corresponded to "Twice a month".

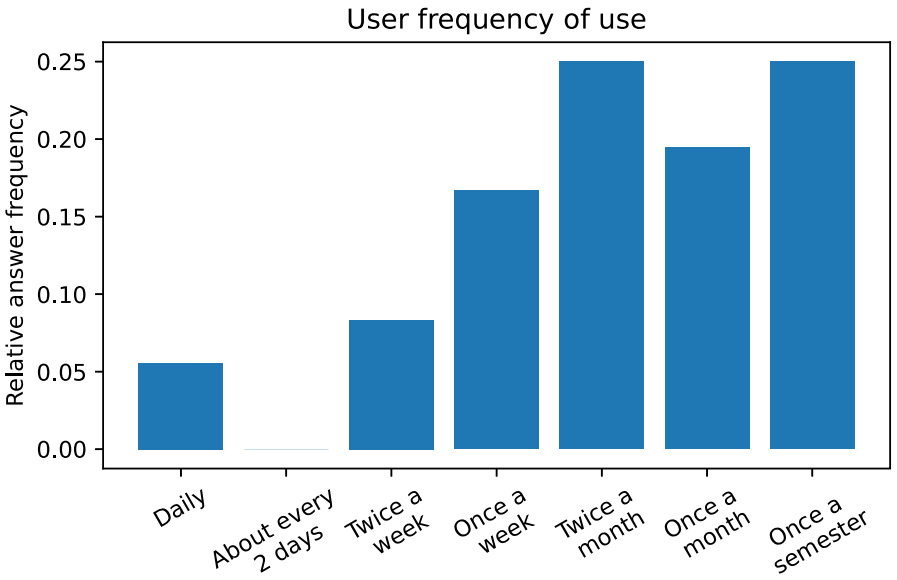


Fig. 6 Frequency of all participants which stated that they used the system more often than "never" and who were therefore categorized as users. The categories have been translated from German to English and their selection frequencies are relative to the number of users

The Likert-type questionnaire contained questions of the different categories “Intention of use”, “Perceived Ease of Use”, “Perceived Usefulness”, “Relevance to the Course of Study” and “Output Quality”. Each of the above categories corresponds to the averaged results of a group of several questions listed in Appendix A. The responses to each question are depicted in Fig. 7. To facilitate comprehension, we consolidated the results of all questions according to their respective categories in Fig. 8. This figure shows the frequencies of the ratings of these questionnaire categories relative to all participants. The numbers in the figure correspond to a scale from 1=“strongly disagree" to 7=“strongly agree", with 4=“neutral". A rating of 7 on the scale, corresponding to “strongly agree”, indicates a positive outcome for RATsApp. The violin plot on the **left** in each category shows the rating of users of the platform, while the violin plot on the **right** shows the ratings of non-users. The plot also includes the frequency of participants who did not provide an answer. In general, most questions were not answered by about 5% to 10% of the users. For simplicity, we interpret a rating of 5 or higher as confirmation of the question asked, since 4 is the neutral rating.

Most of the questions, except the questions for the intention of use, required answers based on actual experience with the system. Many non-users also answered these questions, often choosing the “neutral” answer option. From this point forward, for a more detailed evaluation, we focus our analysis solely on system users, specifically the students who have indicated that they used the system as defined at the beginning of this section. Our findings are derived from individual questions, as illustrated in Fig. 7, which shows the unsummarized answers of the users.

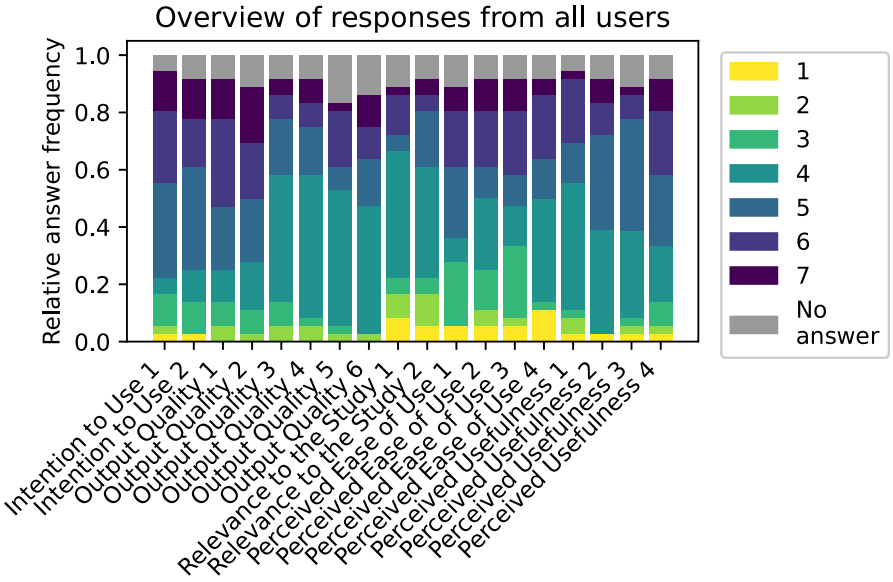


Fig. 7 Answer frequencies of use of participants who stated that they used the system more often than “never" and were therefore categorized as users. The use frequency categories show the selection frequencies relative to the number of users. The questions belonging to each item can be found in Table 3 in Appendix A

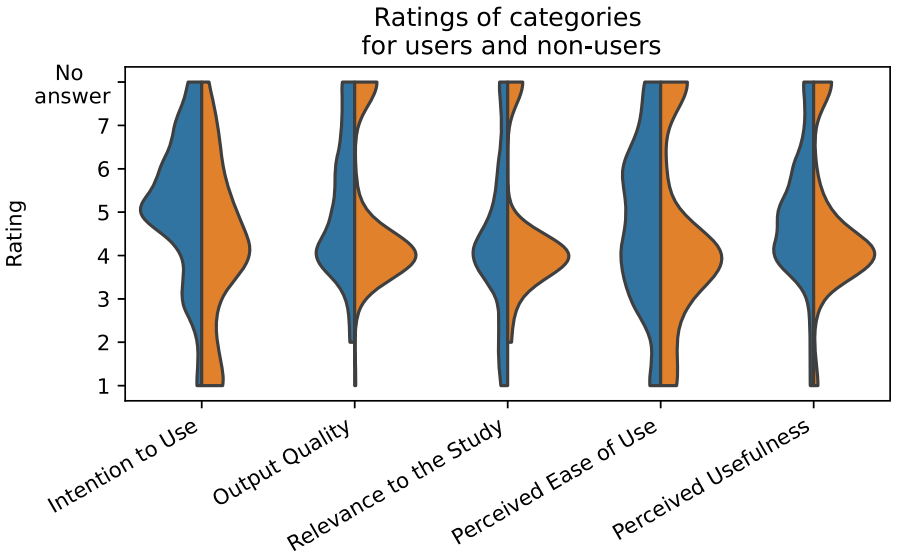


Fig. 8 Distribution of selected rating for a category of the questionnaire relative to the number of all participants. The violin plot on the **left** in each category shows the rating of users of the platform, while the violin plot on the **right** shows the ratings of non-users. The different numbers in the figure correspond to a scale from 1="strongly disagree" to 7="strongly agree" (positive for the platform), with 4="neutral". The plot also includes a representation of participants who did not provide an answer

To analyze our results within the TAM2 usability survey framework (see Section 3.2), we used Pearson correlation analysis to measure the strength and direction of linear relationships between variables, and linear regression analysis to assess the predictive power and significance of specific predictors. These statistical tests were selected to identify significant relationships between the TAM2 question categories, thereby verifying adherence to the TAM2 framework. Figure 9 shows the Pearson correlation values and the significance levels obtained from the linear regression analysis, which will be discussed in more detail below.

Intention to use Around 70% of students rated their intention to use higher than the neutral value of 4. Only about 25% rated it neutral or lower and about 5% did not answer. This is in line with the frequency of use shown in Fig. 6, where over 25%

Fig. 9 Results of the TAM2 analysis: this figure illustrates the interrelationships between the categories using Pearson correlation values. The significance of these linear relations, as determined by linear regression analysis, is indicated as follows: * denotes p <0.05, ** denotes p <0.01, and *** denotes p <0.001

of users indicated that they used RATsApp only once in the semester. It should be noted that despite the fact that many users rated their intention to use as 5 or higher, only around 30% of users used the system weekly or more often. By numbering each answer option for the frequency of use starting with 1 for "daily" and ending with 7 for "Once a semester", we found a significant positive linear relationship ($r = 0.42$, $p = 0.0014$). between the summarized intention of use (shown in Fig. 8) and the frequency of use (shown in Fig. 6). For the non-users, it can be seen that around 35% had an intention of using the system if they were able to.

Output quality Nearly 70% of users considered the quality of the feedback provided after solving each RAT as high, and around 60% of users had no problem with the quality. As a reminder, this feedback was generated by experts in the field. On the other hand, when asked about the quality of the estimation of their competence level, only around 35% of users rated the quality as high and roughly the same amount had no problem with the quality. The quality of the display of the temporal progression of the user's answer correctness was also rated as high by just over 30%. Additionally, around 40% had no problem with its quality.

Our analysis based on TAM2 revealed a significant positive linear relationship between the output quality and the perceived usefulness ($r = 0.34$, $p = 0.011$).

Relevance to the course of study: Only around 20% of the users rated the system as important to their studies and 30% as relevant to their studies. Despite these percentages, our analysis based on TAM2 revealed a significant positive linear relationship between the relevance to the course of study and the perceived usefulness ($r = 0.54$, $p < 0.001$). Interestingly, this relationship was not influenced by the output quality ($p > 0.05$).

Perceived ease of use In terms of the perceived ease of use, around 55% of users felt that interacting with RATsApp was clear and understandable. However, only around 40% of users indicated that they could navigate RATsApp without requiring much effort. Around 45% of users found RATsApp easy to use and only around 40% of users found it easy to make RATsApp do what they wanted to do.

Furthermore, as proposed by TAM2, we observed a significant positive linear relationship between the ease of use and the perceived usefulness ($r = 0.30$, $p = 0.027$). Contrary to the propositions of the TAM2, we did not observe a significant linear relationship between the ease of use and the intention to use ($p > 0.05$). We also looked at how the ease of use of the technology affected the intention to use it, with perceived usefulness acting as an intermediate step in this relationship. In simple terms, we wanted to see if making the technology easier to use would make it seem more useful, and if this increased usefulness would lead to greater intention to use the technology. However, our results showed that ease of use did not have a significant effect on intention to use the technology through perceived usefulness, as the statistical test did not show a significant effect ($p > 0.05$).

Perceived usefulness It followed that around 40% and 50% of the users thought that using RATsApp improved their performance in their studies and that they learned more using RATsApp, respectively. Despite the low percentage of students who found RATsApp relevant to their studies, half of the users tended to think that using RATsApp made them more effective learners and around 60% of users found RATsApp useful for their studies.

In line with the propositions of the TAM2, we found a significant linear relationship between the perceived usefulness and the intention to use ($r = 0.68$, $p < 0.001$).

Finally, we analyzed the written feedback obtained from the survey. The free text questions that were included in the survey are listed in Table 2 of the Appendix A. Based on the users' responses, we created different categories (see Table 1). The categories have been translated from German to English and are listed below:

1. *Cumbersome access to website (VPN, Eduroam, no app):* composed of feedback recommending that RATsApp be developed into an application, be accessible from anywhere and that access in general should be less cumbersome. For example, one study participant commented "Please get rid of the VPN and introduce an app for RATsApp, if possible also offline".
2. *Technical problems/website not working properly:* composed of feedback reporting bugs and other technical problems with the website, most of which were resolved before the survey was conducted. An example response was "Sometimes it did not work as desired (course was not displayed; I could not open RAT sheets on my cell phone)". All issues raised were promptly addressed post feedback.
3. *Missing features in the beginning:* composed of feedback complaining about missing features that were already implemented at the time of the survey, as well as feedback praising features which were introduced between the deployment of the website and the time of the survey. It is unclear why some users reported some features as missing, but it is possible that they were not aware of the recent updates. For example, feedback that fell into this category was "The division into the RAT sheets, like on moodle, helped a lot to repeat specific tasks".
4. *Cumbersome or unclear navigation of the website:* composed of feedback that did not specify further why the navigation was cumbersome as well as specific feedback. An example of such feedback is "The start page could be easier to use and you could go directly to your questions. In addition, it is not always easy to see the competence level".

Table 1 Frequencies of the categories of the received feedback

Category	Frequency
Cumbersome access to website (VPN, Eduroam, no app)	19
Technical problems/website not working properly	7
Missing features in the beginning	6
Cumbersome or unclear navigation of the website	5
RATsApp is useful in principle and a good idea	3
Feedback and graphics from RATsApp are helpful	3
Participation for monetary reasons	2
No time	1
Other	5

5. *RATsApp is useful in principle and a good idea:* composed of feedback that praised RATsApp in general, such as “RATsApp is a very good complementary tool to check and improve your learning progress in a playful way”.
6. *Feedback and graphics from RATsApp are helpful:* composed of feedback that specifically complemented the output of our system, such as “Otherwise, the website also shows quite well how the physics skills develop over the course of the exercises and I find the explanations to the tasks immediately following great”.
7. *Participation for monetary reasons:* Some users wrote that they also participated for monetary reasons. So we put feedback like “The sweepstakes are really great :) Can use the “pocket money" well” in this category.
8. *No time:* Only one participant cited lack of time in the feedback. We still put this in a separate category because it was one of our concerns that students would not have not enough time for RATsApp because of their other duties. The exact feedback was “Unfortunately, I did not find time to use the website besides working on the practice sheets”.
9. *Other:* contains other feedback that was only mentioned once. An example for this is “However, the solution paths are not always easy to follow”.

The frequencies of the free text categories shown in Table 1 provide insight into the user ratings shown in Fig. 8. Difficulty accessing the site, along with technical problems and missing features at the beginning of the semester, likely contributed to the low usage rates of RATsApp, despite the high reported intention to use and perceived usefulness for studying. Another common element of feedback was cumbersome or unclear site navigation, which may explain why less than half of participants rated the system as easy to use, highlighting an area for improvement.

Despite these challenges, users expressed appreciation for the concept behind RATsApp and found its feedback helpful.

6 Discussion

In this section, we analyze and interpret the findings from our usability survey, adhering to the structure delineated by the TAM2 framework, as proposed by Venkatesh and Davis (2000). As mentioned in Section 3.2, this framework was also instrumental in the construction of our usability survey.

Our usability survey revealed a strong inclination towards the use of RATsApp. A substantial majority of participants, constituting 70%, expressed an intention to use it. According to the TAM2 framework, the intention to use a system is a key determinant in whether an individual will adopt or reject a new system. In our results, we identified a significant linear relation between the intention to use and the frequency of use. This finding aligns with the principles of TAM2 and suggests that the lack of an intention to use the system could be a primary factor leading to its non-use. This linear relation highlights the importance of fostering a strong intention to use among potential users.

Therefore, in our discussion, it is crucial to address any factors that may be impeding this intention in order to increase system usage.

Regarding perceived ease of use, consistent with TAM2, we identified a significant linear relationship between ease of use and perceived usefulness. However, diverging from TAM2, we did not find a significant linear relationship between ease of use and intention to use. In Venkatesh and Davis (2000), this relationship was found to have a minor effect, leading us to hypothesize that our sample size may not be large enough to detect this relationship. Similarly, unlike TAM2, we did not find any significant direct or indirect effects when we looked at how easy the system is to use, how useful people find it, and their intention to use it. This could likely be due to the same reason mentioned above.

In terms of perceived usefulness, our findings revealed a significant linear relation between perceived usefulness and intention to use, which is in line with TAM2. We also observed a significant linear relation between output quality and perceived usefulness, as well as a highly significant linear relation between relevance to the course of studies and perceived usefulness. These findings are consistent with TAM2. However, unlike TAM2, this linear relationship was not moderated by output quality.

Additional external factors, such as social influence processes considered by TAM2 but not evaluated in this study, could also contribute to the perceived usefulness. For instance, students might feel compelled to use the website due to external influences or expectations, such as recommendations from teachers or peers, regardless of their personal assessment of its usability or utility. As discussed by Pfeffer (1982), and corroborated by the TAM2 (Venkatesh & Davis, 2000), adapting a certain behavior, like using a system, could enhance a person's standing within the group if influential members of that person's workplace social group believe that it should be adopted. This context highlights the potential limitations of our analysis by acknowledging that factors beyond ease of use and perceived usefulness could influence students' intentions to use the website. We plan to investigate these additional factors in future work.

In light of our analysis, which did not find a significant linear relation between the ease of use and the intention to use RATsApp, it appears that enhancements in the user interface, such as design and layout modifications, may not significantly increase its usage. Despite our results suggesting that it may not be of significant importance, we plan to enhance the user experience of RATsApp to encourage wider adoption and consistent usage. Our primary focus will be on improving navigation and accessibility. We will incorporate user feedback to refine the design and functionality of RATsApp, making it more intuitive and user-friendly.

A more effective strategy for boosting RATsApp usage could be to increase its relevance to the course of study. This could be achieved, for instance, by better aligning the RATs with lecture content and examination material. It is also crucial that the benefits of the system are effectively communicated to users to emphasize its relevance.

Additionally, our results suggest that while less consequential, refining output quality could also contribute positively to its usage. The output quality subcategories

include some of RATsApp's most critical features, making their improvement also a priority. One such feature is the assessment of STEM competency, which includes data literacy, mathematical, and representational competencies. The usability test indicated potential for enhancement, with only 35% of students rating the output quality as high.

A further aspect associated with output quality is the dashboard. Merely 30% of students rated the dashboard of RATsApp highly, indicating a potential area for improvement. The visualization of the graphs in the dashboards may require enhancement to improve readability and comprehension, a critical factor considering that learning dashboards are known to promote awareness, reflection, and sense-making by visualizing learning activity traces.

Moreover, the output quality in RATsApp was found to be significantly influenced by the provision of formative feedback, which was highly valued by approximately 70% of students. Around 60% of students reported no issues with the quality of this feedback, indicating its usefulness and comprehensibility. Formative feedback, essential for enhancing student comprehension and skills (Kulhavy & Stock, 1989), is typically delivered at the lecturer's discretion in traditional classrooms. However, in e-learning environments like RATsApp, the lack of face-to-face interaction poses a challenge for timely and effective feedback. Our system's expert-driven feedback design is based on the theoretical framework of Hattie and Timperley (2007) (see also Wisniewski et al. 2020. This expert-driven approach was observed to be appealing to students in our study. Our findings suggest that this particular aspect of RATsApp was instrumental in positively influencing the perceived usefulness of the system, thereby increasing the students' intention to use it.

Finally, the fact that a third of non-users expressed an intention to use the system, despite not currently doing so, implies that technical and access-related issues are likely the main obstacles to usage. According to TAM2, the high intention to use the system among non-users signals a perceived usefulness suggesting they recognize the website's potential value, despite not experiencing the system first hand. This is another indicator for external factors influencing the results. This emphasizes the necessity to simplify system access and reduce barriers such as VPN requirements. It also highlights the website's potential to draw more users once these external barriers are mitigated.

6.1 Limitations

6.1.1 Website limitations

Despite our findings, the website still has some limitations, which are explained below. One of these limitations is the criteria list to estimate the competence levels lacks generalizability, and therefore it may not accurately identify the particular difficulties a student experiences. For example, a low score in representational competence may be due to difficulties in interpreting a pie chart, but not necessarily in understanding

a graph. In addition, because RATs can assess multiple competencies simultaneously. This means that a single RAT may not discriminate between different competencies, and a lower score in a particular competence after completing the RAT does not clearly delineate the specific competence in which the student is deficient. These factors may distort the estimation of competencies and contribute to the fact that competence levels are only estimates. The calculated level strongly depends on the given RATs and may not reflect a high absolute competence level. These limitations make it important to test how well our system estimates competencies and to have an exercise pool which fulfills diverse criteria of different difficulties. However, this evaluation falls outside the scope of the present manuscript and will be addressed in future work.

6.2 Future work

To address the present limitations of the website, we plan to assign weights to each competence across all criteria in our list to reflect that the importance of each criterion in determining a certain competence may vary, when averaged across criteria. This means that a certain criterion may be more relevant to determine a competence level than another. Moreover, a RAT may draw more heavily on certain competencies than others, so it would be necessary to identify relevance levels for each competence and each RAT. Furthermore, we want to weigh the influence of different RATs on the competence estimation. By giving more difficult RATs higher weights in the computation compared to easier RATs, as in RATs that can be answered correctly with a lower competence level, we aim to increase the precision of our competence estimation. In further studies, we will evaluate which criteria to keep, which to replace and which we need to add to effectively estimate the competence levels. Also, we want to investigate the influence of giving competence estimates as feedback on the final lecture grade.

Moreover, we plan to integrate machine learning (ML) algorithms to predict students' scores in the final exams based on their performance. This could potentially help the lecturer as an alert system to provide additional support to students who need it. We also plan to use ML to suggest RATs to students according to the competence they need to strengthen.

In addition to the planned technical improvements, an ethical focus on positively shaping the practical field of learning and teaching will also continue. It is in our interest to go beyond a purely ethically unproblematic design of the system and data security concerns, and to broaden the view to the actors involved in the learning environment and their relation to each other. As mentioned earlier, we also want to consider the results that are relevant from an ethical perspective. This includes, among other things, the technical optimization of the RATsApp regarding the enhancement of interpersonal discourse. Further development should prioritize interaction between students and/or between students and instructors. One potential approach to facilitate this would be to establish learning groups for students and/or teaching-learning groups where students can engage in collaborative dialogue regarding their thoughts,

challenges, and feedback (Solanki et al., 2019; Wilson et al., 2015; Stadtfeld et al., 2019).

Furthermore, as it is possible that students may be unable to answer RATs beyond their existing learning competencies - whether due to perceived pressure to solve tasks correctly, anxiety related to exams, or lack of self-confidence - it is both ethically and didactically necessary to critically reflect on these aspects. Therefore, if required, it may be helpful to provide or facilitate interpersonal communication options, in addition to technical options and support, helping to overcome these challenges. This could potentially reduce the barriers between students and lecturers, for example, by posing a question anonymously, or by placing service offerings related to exam anxiety or other relevant topics in the RATsApp layout.

It also is imperative to address the potential for failure due to a mismatch between the student's abilities and their chosen career path. In the event that a student consistently provides incorrect responses to tasks, a safety mechanism should be implemented. This could involve providing students with the option to directly consult with a lecturer and/or student counselor for interpersonal feedback and assistance in developing an accurate self-assessment. If this feedback is delivered in an honest and appreciative manner, a potential failure in the examination may present an opportunity for the student to explore ernative educational pathways that better align with their individual strengths and interests.

7 Conclusion

The creation of RATsApp - an open source AFS - is an interdisciplinary effort between physics and mathematics education researchers, engineers, computer scientists, and ethicists. For the development, we have considered several relevant research results on formative feedback, formative assessment, student and teacher dashboards, scaffolding support, STEM competence level estimation, and learning analytics. Given the high demand for personalized learning tools (Harasim, 2000; Anderson, 2003), and considering that STEM competencies are at the forefront of today's challenges (Blumenthal, 1961; Kippers et al., 2018; Küchemann et al., 2021), one must assume that there is a significant demand for personalized learning tools in STEM across different levels of education. This was confirmed in the context of higher education by our usability survey, where around 70% of users rated their intention to use these tools as high.

Our study found that the survey results largely align with the predictions of the TAM2 framework (Venkatesh & Davis, 2000). We observed a significant linear relation between the intention to use and the frequency of usage, emphasizing the importance of fostering usage intention. While ease of use correlates with perceived usefulness, it does not significantly correlate with the intention to use. However, perceived usefulness does significantly correlate with intention to use, output quality, and course of study relevance. This suggests that while certain areas, particularly visual and navigational

elements, could benefit from refinement, interface enhancements may not necessarily lead to a substantial increase in usage. Instead, it might be more effective to align RATsApp with lecture content and exam material to enhance its relevance to the study. Furthermore, improving the output quality, especially for critical features like STEM competency and the dashboard, could have a positive impact on usage.

The survey results are promising overall. Most students responded positively to the expert-generated feedback, and a majority find the system beneficial. The positive feedback from students underscores the promising future of our website application in supporting and facilitating learning. Therefore, addressing the issues mentioned above could further enhance the perceived usefulness and ease of use of RATsApp, thereby maintaining high system usage.

On a final note, it is crucial to understand that this manuscript focuses on a usability survey, which mainly reflects students' perceptions and may not directly quantify the effectiveness of learning. For an accurate assessment of learning progress, a comprehensive study should be conducted. This could involve pre- and post-tests to measure knowledge acquisition, monitoring individual student progress over time, or even incorporating a control group for comparison. Despite these considerations, our innovative approach of integrating formative feedback with STEM competence level estimation has demonstrated potential and is well-equipped to assist students in their academic journey.

We are committed to helping students and teachers in this fast-growing era of technology by creating an accessible tool for education that makes learning more effective and more sustainable. Therefore, we have made RATsApp publicly available and invite contributions to its ongoing development. The source code can be accessed on GitLab: https://gitlab.rhrk.uni-kl.de/ki4tuk/ratsapp. This manuscript describes version 1.0 of the project.

We wish to highlight the critical role that open-source projects like RATsApp play in enhancing education by providing accessible tools for learning that can be freely used, modified, and shared by anyone. Furthermore, RATsApp can also serve as a valuable research tool, enabling researchers to evaluate educational frameworks. Such projects promote collaboration and innovation while ensuring transparency and sustainability.

Appendix A: Usability survey

The usability survey employed to evaluate RATsApp comprised three components: multiple-choice questions, a Likert-scale questionnaire, and written feedback, as detailed in Table 2.

In the main text (see Fig. 8), we presented aggregate results for each category from the Likert scale questionnaire (see Section 5). Table 3 shows these categories and their individual items. The numbers of each item in a category correspond to those shown in Fig. 7.

Table 2 Usability questionnaire used to evaluate RATSAPP

English translation	Input type
I attended the following lecture	Multi select field
I attended the following 2nd lecture [1]	Multi select field
I attended the following 3rd lecture[1]	Multi select field
I attended the following 4th lecture[1]	Multi select field
I attended the following 5th lecture[1]	Multi select field
How old are you?	Number field
What is your gender?	Multi select field
Do you have a RATsApp account?	Multi select field
How often do you use RATsApp?	Multi select field
Do you have any general comments about RATsApp?	Free text field
I plan to use RATsApp, if I can access it.	7 point Likert scale
I think it is likely that I will use RATsApp when I can access it.	7 point Likert scale
Do you have any other comments on your intention of use?	Free text field
Using RATsApp will improve my performance in my studies.	7 point Likert scale
By using RATsApp, I learn more.	7 Point Likert Scale
By using RATsApp, I learn more effectively.	7 point Likert scale
I find RATsApp useful for my studies.	7 point Likert scale
Do you have any other comments about the usefulness of RATsApp?	Free text field
It is clear and understandable how I need to interact with RATsApp.	7 point Likert scale
I do not have to think much about how to use RATsApp.	7 point Likert scale
I find RATsApp is easy to use.	7 point Likert scale
I find it easy to get RATsApp to do what I want.	7 point Likert scale
Do you have any other comments about the usability of RATsApp?	Free text field
Using RATsApp is important for my studies.	7 point Likert scale
The use of RATsApp is relevant to my studies.	7 point Likert scale
Do you have any other comments on the relevance of RATsApp?	Free text field
The quality of the the feedback provided for the different answer options is high.	7 point Likert scale

[1] The question was displayed only if the previous question was answered

Table 2 continued

English translation	Input type
I have no problem with the quality of the feedback provided for the different answer choices.	7 point Likert scale
The quality of the estimation of my competence level is high.	7 point Likert scale
I have no problem with the estimation of my level of competence.	7 point Likert scale
The quality of the display of the temporal progression of my answer correctness is high.	7 point Likert scale
I have no problem with the quality of the display of the temporal progression of my answering skill.	7 point Likert scale
Do you have any other comments about the quality of RATsApp's output?	Free text field

[1] The question was displayed only if the previous question was answered

Table 3 Questions included in the Likert scale questionnaire presented in Section 5

Category	Items
Intention to use	1) I plan to use RATsApp, if I can access it.
	2) I think it is likely that I will use RATsApp when I can access it.
Output quality	1) The quality of the explanations of the different answer options is high.
	2) I have no problem with the quality of the the feedback provided for the different answer choices.
	3) The quality of the estimation of my competence level is high.
	4) I have no problem with the estimation of my level of competence.
	5) The quality of the display of the temporal progression of my answer correctness is high.
	6) I have no problem with the quality of the display of the temporal progression of my answering skill.
Relevance to the course of study	1) Using RATsApp is important for my studies.
	2) The use of RATsApp is relevant to my studies.
Perceived ease of use	1) It is clear and understandable how I need to interact with RATsApp.
	2) I do not have to think much about how to use RATsApp.
	3) I find RATsApp is easy to use.
	4) I find it easy to get RATsApp to do what I want.
Perceived usefulness	1) Using RATsApp will improve my performance in my studies.
	2) By using RATsApp, I learn more.
	3) By using RATsApp, I learn more effectively.
	4) I find RATsApp useful for my studies.

Appendix B Criteria for estimating STEM competencies

Table 4 This table lists the criteria derived from Scheid et al. (2017), OECD (2018), and Schüller et al. (2019) and indicates the associated competencies

Criteria	Data literacy	Rep. competency	Mathematical literacy
It must be assessed whether the given data are complete and free of errors (also with regard to possible erroneous conclusions and the applicability of analysis procedures and algorithms).	x		
An analysis procedure must be applied.	x		
A suitable visualization for the data must be chosen.	x	x	
A conclusion must be reached using (the visual representation of) data (e.g., based on correlations/connections in the data) or a mathematical result.	x		
A decision must be made based on analysis or statistics.	x		
The consequences of the mathematical result for the real problem must be recognized.			x
A suitable mathematical description of the problem must be chosen. Here, mathematical concepts must be taken into account and suitable assumptions must be made.	x	x	x
Mathematical definitions have to be used.			x
Mathematical formalisms have to be used.			x
Mathematical algorithms have to be used.			x
A calculation must be performed.			x
The limitations of the mathematical model used must be recognized.	x		x
The relevant variables in a problem must be identified (necessary, among other things, to simplify a problem).	x		x
A suitable solution strategy must be selected from a list.			x
In addition to the question, there is also task-relevant text.		x	
In addition to the question, there is at least one picture.		x	
In addition to the question, there is at least one diagram (e.g. graphs, tables, vector fields).		x	
In addition to the question, there is at least one formula or mathematical symbol relevant to the task.		x	

Table 4 (Continued)

Criteria	Data literacy	Rep. competency	Mathematical literacy
At least one connection is required between the forms of representation used.		x	
The distribution of information over the involved forms of representation is partly redundant.		x	
At least one used representation has a high degree of abstractness		x	

As outlined in the main text, our objective is to assess three fundamental STEM competencies: mathematical literacy, data literacy, and representational competence. To achieve this, we have developed a list of objective criteria for RATs, as presented in Table 4, each of which is associated with one or more competencies and can be utilized to determine the requisite competence level for each competence category in a RAT. Our list is derived from the criteria established by Scheid et al. (2017), OECD (2018), and Schüller et al. (2019).

Acknowledgements We thank the lecturers who supported the implementation of the system: Dr. Florentine Kämmerer, Dr. Jean-Pierre Stockis, Prof. Dr. Jan von Delft, Prof. Dr. Vladislav Yakovlev, Dr. Ioachim Pupeza, and Dr. Jonathan Bortfeld.

Funding Open Access funding enabled and organized by Projekt DEAL. This work was supported by BMBF within the project KI4TUK.

Availability of Data and Materials The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

References

Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M- A., Gambs, S.. Voarino, N. (2018). *Montréal declaration for a responsible development of artificial intelligence*. Université de Montréal.

Anderson, T. (2003). Modes of interaction in distance education: Recent developments and research questions. *Handbook of Distance Education*, 129–144

Ballard, C. L., & Johnson, M. F. (2004). Basic math skills and performance in an introductory economics class. *The Journal of Economic Education, 35*(1), 3–23. https://doi.org/10.3200/JECE.35.1.3-23

Bangert-Drowns, R. L., Kulik, C.-L.C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61*(2), 213–238. https://doi.org/10.3102/00346543061002213

Bannert, M., & Mengelkamp, C. (2013). Scaffolding hypermedia learning through metacognitive prompts. *International Handbook of Metacognition and Learning Technologies, 171–186*,. https://doi.org/10.1007/978-1-4419-5546-3_12

Beauchamp, T., & Childress, J. (2019). Principles of biomedical ethics: Marking its fortieth anniversary. *The American Journal of Bioethics*, *19*(11), 9–1. https://doi.org/10.1080/15265161.2019.1665402

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Blumenthal, R. H. (1961). Multiple instruction and other factors related to achievement in college physics. *Science Education, 45*(4), 336–342. https://doi.org/10.1002/sce.3730450409

Bowen, G. M., & Roth, W.-M. (1998). Lecturing graphing: What features of lectures contribute to student difficulties in learning to interpret graph? *Research in Science Education, 28*, 77–90. https://doi.org/10.1007/bf02461643

Brookhart, S.M. (2017). *How to give effective feedback to your students*. Ascd.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65*(3), 245–281. https://doi.org/10.3102/00346543065003245

Chandra, S., & Palvia, S. (2021). Online education next wave: Peer to peer learning. *Journal of Information Technology Case and Application Research, 23*(3), 157–172.

Chen, L., Geng, X., Lu, M., Shimada, A., & Yamada, M. (2023). How students use learning analytics dashboards in higher education: A learning performance perspective. *SAGE Open, 13*(3), 21582440231192150. https://doi.org/10.1177/21582440231192151

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education, 162* https://doi.org/10.1016/j.compedu.2020.104094

Duval, E., Klerkx, J., Verbert, K., Nagel, T., Govaerts, S., Parra Chico, G.A.. Vandeputte, B. (2012). Learning dashboards & learnscapes. *Educational Interfaces, Software, and Technology*, 1–5,

Edelsbrunner, P. A., Malone, S., Hofer, S. I., Küchemann, S., Kuhn, J., Schmid, R., & Lichtenberger, A. (2023). The relation of representational competence and conceptual knowledge in female and male undergraduates. *International Journal of STEM Education, 10*(1), 1–19. https://doi.org/10.1186/s40594-023-00435-6

Engel, C., Linhardt, L., & Schubert, M. (2024). Code is law: How compas affects the way the judiciary handles the risk of recidivism. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09389-8

European Commission, Directorate-General for Communications Networks, Content and Technology (2019). *Ethics guidelines for trustworthy AI*. Publications Office. https://data.europa.eu/doi/10.2759/346720

European Commission, Directorate-General for Education, Youth, Sport and Culture (2022). *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/153756

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410–8415. https://doi.org/10.1073/pnas.1319030111

Fuchs, T. (2020). *Verteidigung des menschen: Grundfragen einer verkörperten anthropologie*. Suhrkamp Verlag.

Hake, R.R. (2002). Relationship of individual student normalized learning gains in mechanics with gender, high-school physics, and pretest scores on mathematics and spatial visualization. *Physics education research conference 8*, pp. 1–14

Harasim, L. (2000). Shift happens: Online education as a new paradigm in learning. *The Internet and Higher Education, 3*(1–2), 41–61. https://doi.org/10.1016/S1096-7516(00)00032-4

Hartman, H.J. (2001). *Metacognition in learning and instruction: Theory, research and practice 19* Springer Science & Business Media.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hattie, J., & Jaeger, R. (1998). Assessment and classroom learning: A deductive approach. *Assessment in Education: Principles, Policy & Practice, 5*(1), 111–122. https://doi.org/10.1080/0969595980050107

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hudson, H., & Liberman, D. (1982). The combined effect of mathematics skills and formal operational reasoning on student performance in the general physics course. *American Journal of Physics, 50*(12), 1117–1119. https://doi.org/10.1119/1.12895

Joisten, K. (2004). Topik, kritik und geometrische methode die bedeutung von giambattista vicos "liber metaphysicus." *Deutsche Zeitschrift für Philosophie, 52*(4), 541–552. https://doi.org/10.1524/dzph.2004.52.4.541

Joisten, K. (2006). Die Hauptstraße verlassen. Oder: Mit Giambattista Vico auf einer anderen Fährte. K. Broese, A. Hütig, O. Immel, and R. Reschke (Eds.), *Vernunft der Aufklärung - Aufklärung der Vernunft* (pp. 53–62). Berlin: Akademie Verlag. [2023-03-31] https://doi.org/10.1524/9783050082073.53

Joisten, K. (2020a) Ethik und digitalisierung, oder: Ethik für ki-systeme. eine grundlegung. *Zukunft der pflege, konferenzband zur 3. öffentlichen konferenz des bmbf-clusters: Zukunft der pflege* (pp. 11–14). PPZ Nürnberg.

Joisten, K. (2020b) Kulturphilosophische Zugänge zur Medizin- und Informationstechnik. A. Manzeschke and W. Niederlag (Eds.), *Ethische perspektiven auf biomedizinische technologie* (pp. 91–99). De Gruyter.

Joisten, K., Thiemer, N., Renner, T., Janssen, A., Scheffler, A. (2022). Focusing on the ethical challenges of data breaches and applications. *2022 ieee international conference on assured autonomy (icaa)* (pp. 74–82).

Kaminski, A. (2020). Gründe geben. Maschinelles Lernen als Problem der Moralfähigkeit von Entscheidungen. K. Wiegerling, M. Nerurkar, and C. Wadephul (Eds.), *Datafizierung und big data: Ethische, anthropologische und wissenschaftstheoretische perspektiven* (pp. 151–174). Wiesbaden: Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-27149-7_6

Kemp, P. (1992). *Das unersetzliche: eine technologie-ethik*. Wichern.

Kippers, W. B., Poortman, C. L., Schildkamp, K., & Visscher, A. J. (2018). Data literacy: What do educators learn and struggle with during a data use intervention? *Studies in Educational Evaluation, 56*, 21–31. https://doi.org/10.1016/j.stueduc.2017.11.001

Kluger, A. N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science, 7*(3), 67–72. https://doi.org/10.1111/1467-8721.ep10772989

Knoop-van Campen, C. A., Wise, A., & Molenaar, I. (2021). The equalizing effect of teacher dashboards on feedback in k-12 classrooms. *Interactive Learning Environments, 1–17*,. https://doi.org/10.1080/10494820.2021.1931346

Kohl, P. B., & Finkelstein, N. D. (2006). Effect of instructional environment on physics students' representational skills. *Physical Review Special Topics-Physics Education Research, 2*(1). https://doi.org/10.1103/PhysRevSTPER.2.010102

Korpershoek, H., Kuyper, H., & van der Werf, G. (2015). The relation between students'math and reading ability and their mathematics, physics, and chemistry examination grades in secondary education. *International Journal of Science and Mathematics Education, 13*, 1013–1037. https://doi.org/10.1007/s10763-014-9534-0

Küchemann, S., Malone, S., Edelsbrunner, P., Lichtenberger, A., Stern, E., Schumacher, R., & Kuhn, J. (2021). Inventory for the assessment of representational competence of vector fields. *Physical Review Physics Education Research, 17*(2). https://doi.org/10.1103/PhysRevPhysEducRes.17.020126

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1*, 279–308. https://doi.org/10.1007/BF01320096

Lerner, R. M. (2013). *At the forge: Sqlalchemy. Linux Journal, 2013*(226), 5.

Lin, T- C., Hsu, Y- S., Lin, S-S., Changlai, M-L., Yang, K- Y., Lai, T- L. (2012). A review of empirical evidence on scaffolding for science education. *International Journal of Science and Mathematics Education, 10*, 437–455. https://doi.org/10.1007/s10763-011-9322-Z

Liu, M., & Yu, D. (2022). Towards intelligent e-learning systems. *Education and Information Technologies, 1–32*,. https://doi.org/10.1007/s10639-022-11479-6

Llamas, A., Vila, F., Sanz, A. (2012). Mathematical skills in undergraduate students. A ten-year survey of a plant physiology course. *Bioscience Education*, *19*(1), 1–10https://doi.org/10.11120/beej.2012.19000006

Meyer, W- U., Bachmann, M., Biermann, U., Hempelmann, M., Ploger, F-O., Spiller, H. (1979). The informational value of evaluative behavior: Influences of praise and blame on perceptions of ability. *Journal of Educational Psychology, 71*(2), 259. https://doi.org/10.1037/0022-0663.71.2.259

Mousavinasab, E., Zarifsanaiey, N., Niakan Kalhori, R., & S., Rakhshan, M., Keikha, L., Ghazi Saeedi, M. (2021). Intelligent tutoring systems: A systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments, 29*(1), 142–163. https://doi.org/10.1080/10494820.2018.1558257

Muthuprasad, T., Aiswarya, S., Aditya, K., & Jha, G. K. (2021). Students' perception and preference for online education in India during covid-19 pandemic. *Social Sciences & Humanities open, 3*(1). https://doi.org/10.1016/j.ssaho.2020.100101

Myers, D. S., & Chatlani, N. (2017). Implementing an adaptive tutorial system for coding literacy. *Journal of Computing Sciences in Colleges, 33*(2), 260–267.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. *Handbook of research on educational communications and technology* (pp. 125–143). Routledge.

Nussbaum, M. C. (2009). Creating capabilities: The human development approach and its implementation. *Hypatia, 24*(3), 211–215. https://doi.org/10.1111/j.1527-2001.2009.01053.x

OECD (2018). *Pisa 2022 mathematics framework (draft).* Paris, France. https://pisa2022-maths.oecd.org/files/PISA2022MathematicsFrameworkDraft.pdf (Accessed on 18 December 2022)

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, 422. https://doi.org/10.3389/fpsyg.2017.00422

Pardos, Z.A., Tang, M., Anastasopoulos, I., Sheel, S.K., Zhang, E. (2023). Oatutor: An open-source adaptive tutoring system and curated content library for learning sciences research. *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–17).

Pea, R. D. (2004). The social and technological dimensions of scaffolding and related theoretical concepts for learning, education, and human activity. *The Journal of the Learning Sciences, 13*(3), 423–451. https://doi.org/10.1207/s15327809jls1303_6

Pfeffer, J. (1982). *Organizations and Organization Theory*. Pitman Publishing.

Pozo, S., & Stull, C. A. (2006). Requiring a math skills unit: Results of a randomized experiment. *American Economic Review, 96*(2), 437–441. https://doi.org/10.1257/000282806777212486

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology, 20*(4), 444–450. https://doi.org/10.1006/ceps.1995.1030

Puntambekar, S. (2022). Distributed scaffolding: Scaffolding students in classroom environments. *Educational Psychology Review, 34*(1), 451–472. https://doi.org/10.1007/s10648-021-09636-3

Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., & Soloway, E. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences, 13*(3), 337–386. https://doi.org/10.1207/s15327809jls1303_4

Ryan, T., Henderson, M., Ryan, K., Kennedy, G. (2022). Feedback in higher education: Aligning academic intent and student sensemaking. *Teaching in Higher Education*, *29*(4), 860–875, https://doi.org/10.1080/13562517.2022.2029394

Scheid, J., Müller, A., Hettmannsperger, R., & Kuhn, J. (2017). Erhebung von repräsentationaler kohärenzfähigkeit von schülerinnen und schülern im themenbereich strahlenoptik. *Zeitschrift für Didaktik der Naturwissenschaften, 1*, 181–203. https://doi.org/10.1007/s40573-017-0065-4

Schüller, K., Busch, P., & Hindinger, C. (2019). Future skills: Ein framework für data literacy. *Hochschulforum Digitalisierung, 46*, 1–128.

Sen, A. (1987). *On ethics and economics*. Oxford University Press.

Sen, A. (1997). *Resources, values and development*. Harvard University Press.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Solanki, S., McPartlan, P., Xu, D., & Sato, B. K. (2019). Success with ease: Who benefits from a stem learning community? *PloS One, 14*(3). https://doi.org/10.1371/journal.pone.0213827

Stadtfeld, C., Vörös, A., Elmer, T., Boda, Z., & Raabe, I. J. (2019). Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences, 116*(3), 792–797. https://doi.org/10.1073/pnas.1811388115

Steinert, S., Avila, K.E., Ruzika, S., Kuhn, J., Küchemann, S. (2023). Harnessing large language models to enhance self-regulated learning via formative feedback. *arXiv preprint*, arXiv:2311.13984,

Systems and Software Engineering Standards Committee, & Olszewska, J.I. (2021). *Ieee standard model process for addressing ethical concerns during system design: Ieee standard 7000-2021*. United States: IEEE. https://doi.org/10.1109/IEEESTD.2021.9536679

Toyokawa, Y., Majumdar, R., Kondo, T., Horikoshi, I., & Ogata, H. (2023). Active reading dashboard in a learning analytics enhanced language-learning environment: Effects on learning behavior and performance. *Journal of Computers in Education, 1–28*,. https://doi.org/10.1007/s40692-023-00267-x

Van de Pol, J., Volman, M., Oort, F., & Beishuizen, J. (2015). The effects of scaffolding in the classroom: Support contingency and student independent working time in relation to student achievement, task effort and appreciation of support. *Instructional Science, 43*, 615–641. https://doi.org/10.1007/s11251-015-9351-z

Van Leeuwen, A., Janssen, J., Erkens, G., & Brekelmans, M. (2014). Supporting teachers in guiding collaborating students: Effects of learning analytics in cscl. *Computers & Education, 79*, 28–39. https://doi.org/10.1016/j.compedu.2014.07.007

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

Verbert, K., Duval, E., Klerkx, J., Govaerts, S., & Santos, J. L. (2013). Learning analytics dashboard applications. *American Behavioral Scientist, 57*(10), 1500–1509. https://doi.org/10.1177/0002764213479363

Verbert, K., Govaerts, S., Duval, E., Santos, J. L., Van Assche, F., Parra, G., & Klerkx, J. (2014). Learning dashboards: An overview and future research opportunities. *Personal and Ubiquitous Computing, 18*, 1499–1514. https://doi.org/10.1007/s00779-013-0751-2

Wiegerling, K. (2021). Exposition einer theorie der widerständigkeit. *Filozofija i društvo, 32*(4), 641–661.

Wilkinson, S.S. (1980). *The relationship of teacher praise and student achievement: A meta-analysis of selected research*. University of Florida.

Wilson, D., Jones, D., Bocell, F., Crawford, J., Kim, M. J., Veilleux, N., & Plett, M. (2015). Belonging and academic engagement among undergraduate stem students: A multi-institutional study. *Research in Higher Education, 56*, 750–776. https://doi.org/10.1007/s11162-015-9367-x

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087

Wu, H.-K., & Puntambekar, S. (2012). Pedagogical affordances of multiple external representations in scientific processes. *Journal of Science Education and Technology, 21*, 754–767. https://doi.org/10.1007/s10956-011-9363-7

Zichy, M., & Alber, V.K. (2017). *Menschenbilder: Eine grundlegung*. Nomos Verlagsgesellschaft.

Authors and Affiliations

Steffen Steinert[1] · Lars Krupp[2,3] · Karina E. Avila[4] · Anke S. Janssen[5] · Verena Ruf[1] · David Dzsotjan[1] · Christian De Schryver[6] · Jakob Karolus[2,3] · Stefan Ruzika[4] · Karen Joisten[5] · Paul Lukowicz[2,3] · Jochen Kuhn[1] · Norbert Wehn[6] · Stefan Küchemann[1]

✉ Steffen Steinert
steinert.steffen@physik.uni-muenchen.de

✉ Lars Krupp
lars.krupp@dfki.de

✉ Karina E. Avila
kavila@rhrk.uni-kl.de

1 Faculty of Physics, Chair of Physics Education, Ludwig-Maximilians-University Munich, Munich 80539, Germany

2 Embedded Intelligence, German Research Center for Artificial Intelligence, Kaiserslautern 67663, Germany

3 Department of Computer Science, RPTU Kaiserslautern-Landau, Kaiserslautern 67663, Germany

4 Department of Mathematics, RPTU Kaiserslautern-Landau, Kaiserslautern 67663, Germany

5 Department of Philosophy, RPTU Kaiserslautern-Landau, Kaiserslautern 67663, Germany

6 Department of Electrical and Computer Engineering, RPTU Kaiserslautern-Landau, Kaiserslautern 67663, Germany