International Journal of Artificial Intelligence in Education
https://doi.org/10.1007/s40593-025-00486-6

ARTICLE

The Impact of Explainable AI on Teachers' Trust and Acceptance of AI EdTech Recommendations: The Power of Domain-specific Explanations

Yael Feldman-Maggor[1] · Mutlu Cukurova[2] · Carmel Kent[3] · Giora Alexandron[1]

Accepted: 16 May 2025

Abstract

Trust is crucial for teachers' adoption of AI-enhanced educational technologies (AI-EdTech), yet how this trust is formed and maintained remains poorly understood. An aspect of the system design that seems profoundly related to trust is transparency, which can be achieved through explainable AI (XAI) approaches. The present study seeks to explore the dynamic nature of teachers' trust in AI EdTech systems, how it relates to understandability, and XAI's role in enhancing it. Building upon Hoff and Bashir's 'trust in automation' model (2015), we propose a theoretical model that connects these factors. We validated the applicability of the proposed model to AI in Education context using a mixed-method, within-subject design that measured understandability, trust, and acceptance of AI recommendations among 41 in-service chemistry teachers. The results showed a significant positive correlation between the three factors, as anticipated by the model, and demonstrated the heterogeneous understandability of different XAI schemes, with domain-driven schemes superior to data-driven ones. In addition, the study reveals two additional factors influencing teachers' adoption of AI-EdTech: pedagogical perspectives and workload reduction potential. The study provides a theoretical explanation of how different XAI schemes impact trust through understandability. Furthermore, it emphasizes the need for greater attention to XAI, which fosters trust and facilitates the acceptance of AI-EdTech.

Keywords Explainable AI (XAI) · Trust · Acceptance of AI · Understandability

Introduction

AI-enhanced educational technologies (AI Ed-Tech) that provide machine learning insights and recommendations can potentially support teachers in various ways (Nazaretsky et al., 2022a; Siemens, 2013). However, despite the touted potential of

Extended author information available on the last page of the article

Published online: 16 June 2025

AI Ed-Tech to help education make a digital transformation, particularly following the recent introduction of generative AI (Saif et al., 2024), recent research highlights the key role that psychological factors, such as trust, play in influencing teachers' adoption of AI-EdTech (Celik, 2023; Cukurova et al., 2020, 2023; Nazaretsky et al., 2022a). Put simply, teachers may be reluctant to adopt an AI recommendation system because they do not always trust the suggestions it provides (Qin et al., 2020). One reason behind this lack of trust is that AI is often experienced as a "black box" by its users, who might struggle to understand how and why certain algorithms reach specific results and generate particular outputs (Rudin, 2019). Trust is fundamental for teachers to embrace AI-EdTech, as highlighted, for example, by Choi et al. (2023), Khosravi et al. (2022), and Nazaretsky et al. (2022b). However, a review of the literature on the topic reveals that the dynamics of how trust in the recommendations of a specific AI-EdTech system is developed and sustained are not well understood. Several studies found that providing explanations about the machine learning rationale underlying an AI-EdTech system (e.g., Nazaretsky et al., 2022c) or adding explainable AI (XAI) features that explain its decisions (e.g., Guleria & Sood, 2023; Wang et al., 2024), can increase users' trust in the system's recommendation. However, there is no theoretical framework that connects between enhancing educational users' understanding of the process and its output and the users' trust in it. In fact, there is also evidence that being knowledgeable about AI does not necessarily translate to ascribing credibility to AI-EdTech (Cukurova et al., 2020). To close this gap, we propose a theoretical model that suggests how XAI, understandability of the system's recommendations, the formation of trust in them, and their subsequent acceptance are connected. Our model is based on Hoff & Bashir's (2013, 2015) model of 'trust in automation', adapted to the AI-EdTech domain. The adapted model explains how effective XAI schemes that foster understandability increase trust and acceptance and shed light on the dynamic nature of trust.

To study this model, we conducted empirical research with 41 in-service chemistry teachers who used an AI-powered recommendation tool that was developed in a previous study (Nazaretsky et al., 2022c). Trust in AI plays a crucial role in shaping the interactions between various educational stakeholders (Ifenthaler et al., 2024; Nazaretsky et al., 2025; Schiff, 2022). In this study, we focus on teachers. This focus is critical because educators play a central role in shaping educational practices, selecting pedagogical approaches and technologies, and evaluating students' work (Wang et al., 2024; Zawacki-Richter et al., 2019).

The AI tool includes a dashboard enabling teachers to analyze students' performance and assign follow-up activities based on its recommendations. In the experiment, the teachers followed a protocol in which they were requested to conduct an authentic data analysis task with the tool. We evaluated the dynamic nature of trust by examining the change in teachers' trust after they were gradually provided with two types of XAI explanations: feature importance, which we refer to as 'data-driven,' and semantic explanations articulated in curricular terminology that 'speaks' the teachers' pedagogical language, which we refer to as 'domain-driven.'

Our findings showed that understandability, trust, and acceptance of AI-EdTech recommendations are positively correlated. They also indicated that the level of understandability which then impacts trust and acceptance through it, can vary

depending on the type of XAI, with domain-driven explanations fostering greater understandability than data-driven ones. In addition to the pre-defined focus on the connection between XAI, acceptance, and trust, bottom-up analysis of teachers' protocols revealed additional factors that impact acceptance and are unrelated directly to XAI and trust. These factors are also reported, with a cautionary note that they are based on exploratory, qualitative analysis.

The contribution of this work is threefold. First, it suggests a model that explains the relationship between XAI, understandability, trust, and acceptance of AI-EdTech recommendations. Second, it sheds light on the *dynamic* nature of teachers' trust in AI-EdTech and how it may be shifted by appropriate XAI design that communicates in domain terminology that teachers can interpret and apply. Third, it reveals two additional factors influencing teachers' acceptance of AI-EdTech: pedagogical perspectives and workload reduction potential. By that, the research extends our theoretical understanding of trust in AI-EdTech and provides practical insights that can inform the design of trustworthy AI-EdTech that educators are more likely to adopt.

The following sections begin with a theoretical background on trust, acceptance, understandability, and explainability. Subsequently, we present our hypotheses and research questions in detail, basing them on these theoretical backgrounds. Next, we describe our methodology, followed by results and discussion.

Theoretical Background

Trust in AI

Trust is defined as "an attitude of confident expectation in a situation of risk that one's vulnerabilities will not be exploited" (Corritore et al., 2003). In the context of automation, it can be defined as the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004). With the advent of AI as a specialized form of automated system, attention to trust in AI has grown. This shift considers various factors: analyses of potentially catastrophic risks associated with AI (Yudkowsky, 2023), concrete evidence of AI's capability for error (Williams & Yampolskiy, 2021), and the subjective phenomenon of AI Anxiety (Johnson & Verdicchio, 2017). For an overview of research on trust in AI, see Lukyanenko et al. (2022). Eventually, trust in an AI system can determine one's positive or negative perceptions of it (Mohseni et al., 2021) and if and how it will be used (Hancock et al., 2011; Sethumadhavan, 2019). As AI is increasingly integrated into decision-making processes (Jarrahi, 2018), one's trust, or the lack of it, in AI (whether it is justified or not) may ultimately have a large impact on the quality of the decisions taken (Lukyanenko et al., 2022). Misplaced trust in agent systems can result in disastrous consequences (Liu et al., 2022; Quinn et al., 2021), and conversely, inappropriate distrust can lead to the disuse of automated systems (Huang & Bashir, 2017).

It is important to note that trust (in automation) is not a constant construct but rather a dynamic state that can evolve (Lukyanenko et al., 2022; Lumineau & Schilke, 2020). Trust is influenced by numerous factors, including human ones,

the performance of automated systems, and the environment (Hancock et al., 2011; Huang & Bashir, 2017), and can increase, decrease, be repaired, or be maintained (Hoff & Bashir, 2015; Vereschak et al., 2021). Thus, users may have different feelings of trust and mistrust during various stages of their experience with any given system (Mohseni et al., 2021). Understanding the dynamic nature of trust in AI-EdTech and its relation to personal, contextual, and system-related factors remains a major open research gap (Lukyanenko et al., 2022; Stackpole, 2019).

The theoretical model proposed by Hoff and Bashir (2013, 2015) provides a conceptual framework to investigate the factors that influence trust in automation, which are applicable to trust in AI (see Fig. 1). They distinguish between the three types of trust: dispositional, situational, and learned. Dispositional trust represents the variability of individuals' instinctive tendencies to trust the trustee (e.g., automation) and cannot change in the short term. Culture, age, gender, and personality characteristics are the primary sources of variability in this most basic layer of trust. Situational trust varies depending on the specific context of an interaction. Learned trust is the most dynamic aspect of trust and is influenced by a user's previous experiences with a particular system. It varies according to the unique characteristics of that system. Thus, it depends on design features and system performance. Design features are significant because they can impact the user's subjective evaluation of the system's performance. During the course of one interaction, an automated system may perform variably, and its user's trust will likely vary to correspond with the system's real-time performance (Hoff & Bashir, 2013). The dynamic nature of trust is thus a function of user-system interaction. When the performance of an automated system impacts its user's trust, the user's reliance strategy may change. In turn, the extent to which a system is relied upon can affect its performance, thus completing the cycle. To facilitate appropriate trust in automation, designers must

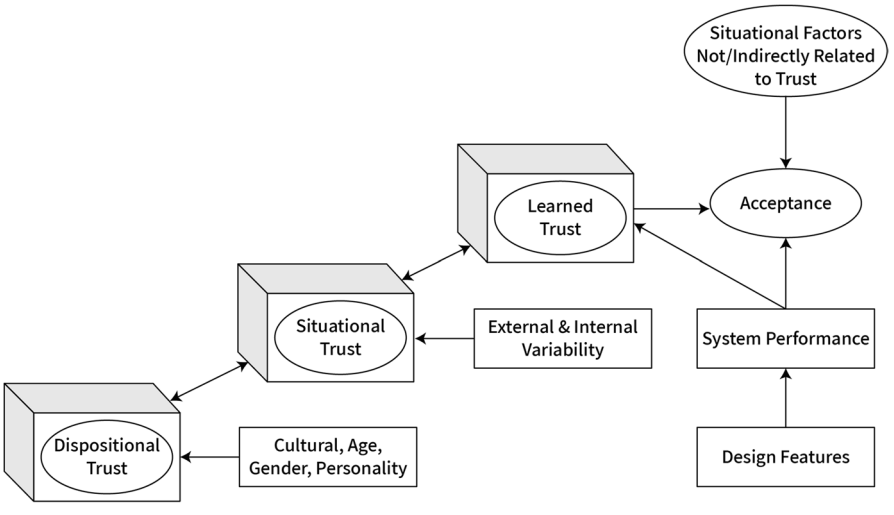

Fig. 1 Model of factors that influence trust in automation (authors' elaboration of Hoff & Bashir, 2015)

carefully consider the interface's ease of use, transparency, and appearance (Hoff & Bashir., 2013). This is in line with the research on trust in AI that highlights transparency in design, which may foster a better understanding of the system and its fair and accurate extent (Kizilcec, 2016), and guides that highlight the importance of transparency in promoting acceptance of AI (OECD, 2021).

Trust in AI EdTech

Trust in AI is important both for students' and teachers' interactions with AI. It is thus becoming an increasingly prominent research topic within the field of AI in Education, primarily examined through the lens of acceptance of technology recommendations among teachers (e.g., Antonietti et al., 2022; Cukurova et al., 2023; Feldman-Maggor et al., 2024) and learners (e.g., Choung et al., 2023; Conijn et al., 2023; Kizilcec, 2016; Maheshwari, 2023).

In the context of students, Nazaretsky et al. (2025) investigated students' trust in AI and found that student demographics, specifically gender and educational background, significantly correlated with their trust perceptions. Expanding the stakeholder scope, Qin et al. (2020) examined trust in AI-based educational systems among students, teachers, and parents through online interviews. Their findings revealed that some parents worry that AI EdTech systems might make students overly dependent on technology, reducing their independent thinking skills, which contributes to parents' reluctance to place continuous trust in these systems. At the same time, many parents expressed concerns about discrimination from teachers. However, AI systems were seen as capable of diagnosing students' needs and providing personalized suggestions based solely on relevant characteristics, thereby treating all students equally and reducing bias.

In the context of teachers, attention was given to characteristics of the human agents' such as cultural differences (Viberg et al., 2024), pedagogical beliefs (Choi et al., 2023), and teachers' understanding of the AI's decision-making processes (Nazaretsky et al., 2022c). The observation that trust in AI is dynamic – namely, it can be learned or temporarily increased – was discussed in previous AI in Education studies (e.g., Wang et al., 2024). Wang et al. (2024) employed explainable AI to clarify the outputs of deep learning models used for classroom dialogue analysis and experimented to evaluate the impact of these explanations on teachers. In their study, fifty-nine pre-service teachers were recruited and randomly assigned to either a treatment or control group. Initially, both groups learned to analyze classroom dialogue using AI-powered models without explanations. Subsequently, the treatment group received both AI analysis and accompanying explanations, while the control group continued to receive only AI predictions. The results showed that teachers in the treatment group demonstrated significantly higher levels of trust in and acceptance of AI-powered models for classroom dialogue analysis compared to those in the control group. Williamson et al. (2025) also found that the explanation raised trust in the AI model.

Explainable AI (XAI) and User Understandability

One of the main ways to enable transparency is by developing AI models that can be explained in layman's terms. In the literature, this concept has been referred to as XAI (Gilpin et al., 2018). More broadly, explainability encompasses everything that makes machine learning models transparent and understandable, including information about the data, performance, and more (Jang et al., 2022; Liao et al., 2020). XAI makes the internal system more transparent, providing explanations of its decisions in a certain level of detail. These explanations are essential to ensure algorithmic fairness, identify potential biases or problems in the training data, and ensure that the algorithms perform as expected (Gilpin et al., 2018).

However, precisely what kinds of explanations are human-interpretable is a complex question (Narayanan et al., 2018). For example, Bussone et al. (2015) found that overly detailed explanations from clinical decision support systems enhance trust yet create over-reliance, while short or absent explanations prevent over-reliance but decrease trust. One study of XAI in education tested three consecutive levels of transparency and found that designing for trust requires balanced interface transparency, not too little, not too much (Kizilcec, 2016). Still, this definition of transparency is qualitative and difficult to replicate. Kizilcec (2016) frames transparent information in a context of judicial fairness and shows that this framed information leads to further adoption.

The main challenge in explainable AI is creating complete and interpretable explanations simultaneously, as these tend to be competing characteristics. This is because there is a tension between the most potent models to deal with complex and non-linear data (e.g., deep learning), which are more difficult to explain, and simpler models that are more explainable (Rudin, 2019), yet often fall short in their prediction power. Another challenge is that XAI developers often express an algorithm-centric view, relying on their own intuition regarding what constitutes a satisfying explanation (Miller, 2019). However, explainability and transparency are subject-dependent factors in that what is transparent to an AI developer might not be transparent to a particular group of end users (Chaudhry et al., 2022). In addition, although explanations can improve users' understanding of AI systems, conclusions about their benefits for user trust and acceptance can be mixed, suggesting potential gaps between algorithmic illustrations and end-user needs (Liao et al., 2020). To address this gap, Liao et al. (2020) developed an algorithm-informed XAI question bank in which user needs for explainability are represented as prototypical questions users might ask about the AI. Their research reviewed four methods of explainability: (1) Explain the model: describe the weights of features used by the model (including visualization that shows the weights of features); (2) Explain a prediction: show how features of the instance contribute to the model's prediction; (3) Inspect counterfactuals: show how the prediction changes corresponding to changes in a feature (often in a visualization format) and describe the feature(s) that will change the prediction if perturbed, absent, or present. (4) Case-based: provide an example(s) with minor differences from the explained instance with a similar prediction.

Understanding and articulating the workings of AI is crucial for both its creators and users (Mohseni et al., 2021). However, different research communities focus on

different aspects of XAI. Research in machine learning seeks to design new interpretable models and provide ad-hoc explanations for black-box models. In contrast, in human–computer interaction (HCI), the focus is mainly on end-user needs such as trust and understanding of machine-generated explanations (Mohseni et al., 2021). For example, Wang and Yin (2022) study the effectiveness of four XAI methods that support users in making better decisions. Their results demonstrate that the XAI methods are only effective in cases where the users have previous knowledge in the relevant domain.

Explanations in AI are often categorized as either *global* or *local* (Khosravi et al., 2022; Mohseni et al., 2021; Setzu et al., 2021). A global explanation offers an overarching view of the ML model's functioning. Techniques like model visualization and decision rules are types of global explanations (Mohseni et al., 2021). Another approach of global explanation is feature importance, which involves assigning and comparing scores to each feature during prediction, indicating their respective importance in the model's output. These scores, which can be calculated using methods like linear regression, logistic regression, or more contemporary approaches like Shapley values (Fréchette et al., 2016), can be communicated through reports or visual graphs (Khosravi et al., 2022).

In contrast, local explanations focus on defining the connections between specific input–output pairs. They clarify the rationale behind the outcomes of individual queries or particular input instances (Mohseni et al., 2021). The need to provide explanations is highlighted in various educational contexts. Teachers require them to ensure accountability when offering personalized feedback to students, to aid teachers in diagnosing areas where a student group may need more attention, and during consultations with parents, assisting them in supporting their child's learning process (Khosravi et al., 2022). However, Conijn et al. (2023) studied students' trust and motivation when using an automated essay scoring system. The study provided students with two explanations: full-text global explanations and an accuracy statement. Their results showed that neither explanation significantly affected students' trust or motivation compared to receiving no explanation at all.

Dikmen and Burns (2022) emphasize an additional aspect of XAI: the significance of domain-specificity when utilizing XAI systems in complex decision-making scenarios. They underline this knowledge's critical role in fostering trust and reliance within the context of XAI. This perspective is also found in the work of Donadello and Dragoni (2020), who demonstrate how semantic-based explainable AI enhances transparency by generating explanations that are understandable to humans. Likewise, Panigutti et al. (2020) illustrate that leveraging semantic information within medical ontologies can significantly improve the quality of explanations.

Additionally, Shin (2021) suggests that an alternative approach to achieving model or algorithm explainability is through understandability, which involves how users interpret algorithmic features and comprehend algorithm-based systems. Addressing these aspects of user interaction is vital as AI becomes increasingly widespread. Understandability is defined as a product of both transparency and interpretability. AI must be valid, reliable, and understandable for it to be considered trustworthy. In this context, understandability, serving as an operational proxy for explainability, requires an AI system to be transparent and interpretable (Joyce et al.,

2023). The idea is that humans should know why they trust an AI tool (Merritt et al., 2013). For instance, Joyce et al. (2023) studied the healthcare sector and described understandability as a proxy of transparency and interpretability. This combination depends significantly on the human operator's ability to align the behavior of algorithms with their professional expertise and knowledge. In other words, understandability can be assessed as a perceived attribute, where users evaluate the quality of explanations based on their own understanding and interpretability levels (Samek et al., 2017; Shin, 2021). We adopt this definition for the present study.

To conclude, the body of AIED research on XAI and Trust lacks a theory that connects these aspects and an explanation of how and why specific XAI schemes contribute to trust and acceptance. The current research takes a step towards closing this gap by adapting Hoff and Bashir's (2013, 2015) 'trust in automation' model. This adapted model places understandability as the factor connecting XAI and trust, and through this, it also explains why the understandability of certain XAI schemes influences their impact on trust.

Research Hypothesis and Research Questions

Based on the adapted model (see below), our hypothesis is that effective XAI interfaces enhance the understandability of the logic behind AI-generated recommendations, supporting the formation of trust in them and, through this, increasing their acceptance. In particular, we are interested in the projection of this hypothesis to AIED contexts and teacher: AI partnership.

The theoretical model that underlies our hypothesis is based on the model initially proposed by Hoff and Bashir (2013; extended significantly in 2015; see Sub-Section "Trust in AI"). This model emphasizes the significance of design features in shaping users' subjective system performance evaluation and how this evaluation enables them to establish and sustain trust in the system's output. By that, the model connects system design, performance evaluation, and dynamic trust. We adapt this model and propose an appropriation based on the following observations. First, that XAI features are part of the system design features. Second, to assess system performance effectively, users need to *understand* its performance (Hoff & Bashir, 2015). Therefore, we argue that understandability is a characteristic of evaluating system performance. The proposed integration of XAI and understandability into the original model is shown in Fig. 2.

Based on the adapted model, the relationship between XAI and dynamic trust becomes clear: XAI increases the system's understandability by enabling the user to interpret its output and validate it against the expected outcome. This validation of the system's performance can establish trust in its behavior.

The study presented here has four primary objectives. The first is validating the relationships among understandability, trust, and acceptance, as Hoff and Bashir's (2013) model suggested. This involves examining the correlations between these factors. The second objective is to investigate the dynamic nature that the model attributes to these constructs. The third objective is to evaluate how understandability and, through this, their trust in the AI-EdTech system and their recommendations

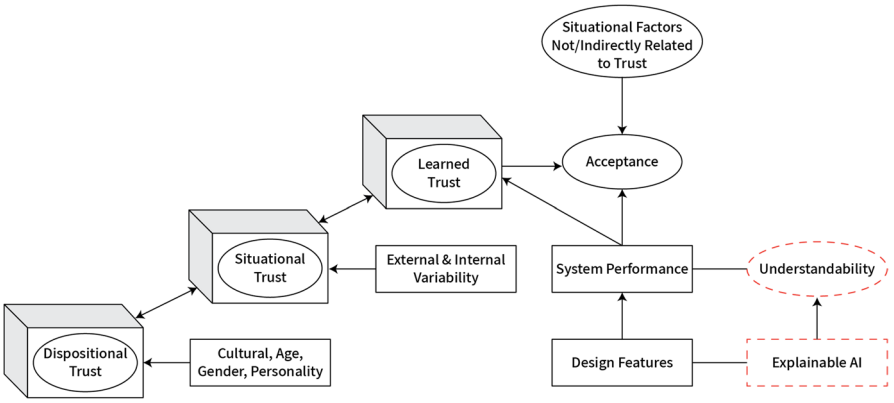

Fig. 2 Model of factors influencing trust in automation, integrating explainable AI and understandability. The new components are dashed

shift after receiving domain-driven explanations following data-driven ones. While data-driven XAI received more attention within the AIED space (e.g., Swamy et al., 2023), building on results outside the education domain (Donadello & Dragoni, 2020; Panigutti et al., 2020), we hypothesize that domain-driven information can significantly enhance the quality of explanations compared to relying solely on data-driven ones. The last objective is to collect, in a bottom-up, qualitative fashion, additional factors that may shape teachers' trust and acceptance of AI tools. These objectives are realised through the following research questions.

RQ1. What is the relationship between understandability, trust, and acceptance, and do these variables change during the teacher's interaction with the system? (Objective 1, 2)
RQ2. To what extent do teachers' perceptions of system performance, trust, and acceptance of AI tools shift after receiving domain-driven explanations following data-driven explanations? (Objective 3)
RQ3. Which additional factors may influence teachers' willingness to accept and rely on AI analysis? (Objective 4)

Methodology

To address the study questions, we conducted a within-subject staged experiment in which teachers performed an AI-aided student data analysis task in which explainability information was gradually unveiled in two conditions. The conditions were data-driven and domain-driven. The research combined qualitative and quantitative tools (mixed methods) that are known to increase the precision and trustworthiness of the results (Leech & Onwuegbuzie, 2007), as described below. All tools described below were validated by the research team and four additional science

education experts who were not part of the research team. The aim of this validation was to ensure that the protocol and items were clear and that they measured what we intended to measure (Torrecilla-Salinas et al., 2019).

The rationale for choosing a within-subject design in this study stems from our aim to evaluate the adapted Hoff and Bashir (2013, 2015) model in the AIED context, which theoretically evaluates changes in individual attitudes. Our hypothesis is that effective XAI interfaces enhance the understandability of the logic behind AI-generated recommendations, thereby supporting the formation of trust and, in turn, increasing their acceptance. A within-subject design allows us to assess this process by demonstrating individual change (Charness et al., 2012).

Participants

The research population included forty-one in-service high-school chemistry teachers (gender makeup: 37 females, four males). The sample displays diverse ages, educational backgrounds, and professional experiences. The teachers completed the research protocol (see Section "The AI Tool, the Procedure, and the Explainability Conditions"), which required 30 to 45 min. Eleven of the 41 teachers participated in a semi-structured think-aloud protocol while conducting the survey. Teachers' teaching and demographic backgrounds appear in Table 1.

The AI Tool, the Procedure, and the Explainability Conditions

The AI Tool

The AI tool that served as the research vehicle was GrouPer – an AI-based recommendation tool developed in previous research (Nazaretsky et al., 2022a). The tool is an AI-for-Teacher technology that supports personalized instruction in the context of science education. Its analysis engine uses an unsupervised machine learning algorithm (cluster analysis) to perform a multidimensional analysis of student responses to interactive assessment items. The clustering method assumes that students perform similarly on items requiring the same skills and competencies (Nazaretsky et al., 2019). It then divides students into groups with similar knowledge profiles ('clusters'). The validity of the tool's underlying algorithms and its ecological validity to in-class formative assessment scenarios that are similar to the ones led by the teachers who participated in the current study were established in previous studies (Din et al., 2023; Nazaretsky et al., 2022a). The clusters are presented in an interactive dashboard that enables teachers to examine each cluster's performance and assign learning activities adapted to the needs of student groups based on their strengths and weaknesses. It was co-designed with teachers in a process described by (Nazaretsky et al., 2021) and is presented in Fig. 3.

As the tool's dashboard proposes a recommendation based on the analysis conducted by the underlying AI engine, it provides an authentic context to study issues of trust and explainability in AI decision-making. Specifically, clustering algorithms generate bottom-up, data-driven grouping based on similarity rather than on

Table 1 Teachers' teaching experience and demographic background

Teachers' Characteristic	Frequency and Percentage (survey only) ($N = 30$)	Frequency and Percentage (Survey + semi think aloud ($N =$ 11)
Age		
Below 40	13 (43.3%)	NA
41–50	9 (30.0%)	8 (72.7%)
Above 50	8 (26.7%)	3 (27.3%)
Experience in teaching (years)		
0–5	8 (26.7%)	NA
6–10	11 (36.6%)	6 (54.5%)
11–20	8 (26.7%)	1 (9.1%)
Above 20	3 (10.0%)	4 (36.4%)
Experience in educational technology		
0–5	13 (43.3%)	2 (18.2%)
6–10	12 (40.0%)	5 (45.4%)
11–20	2 (6.7%)	4 (36.4%)
Above 20	1 (3.3%)	NA
Missing	2 (6.7%)	NA
Educational Background		
Bachelor's Degree	14 (46.7%)	1 (9.1%)
Master's Degree	12 (40.0%)	6 (54.5%)
PhD8 (19%)	4 (13.3%)	4 (36.4%)

predetermined categories. Without predefined categories, the criteria for similarity can often be difficult to comprehend and explain (Moshkovitz et al., 2020).

Introducing the AI-EdTech System to the Teachers

The procedure followed a staged research within-subject design. The teachers did not receive a formal training process, as we aimed to avoid any influence on their trust in the system. However, before the explanation phase, they were provided background information about the GrouPer tool (see Appendix A). In the initial phase, the teachers were introduced to GrouPer and how it operates (This was before presenting the XAI conditions). Teachers were given instructions about the basic dashboard shown in Fig. 3: GrouPer is an AI-powered tool utilizing large-scale data, applying a machine learning algorithm to divide students into groups based on the similarity in their response patterns in a specific assignment. After analyzing many student responses, the algorithm forms groups and assigns each student to their most appropriate group. The circles in the dashboard represent students from the teacher's own class. It is important to note that the clusters presented do not reflect the students' general abilities (like 'excellent' or

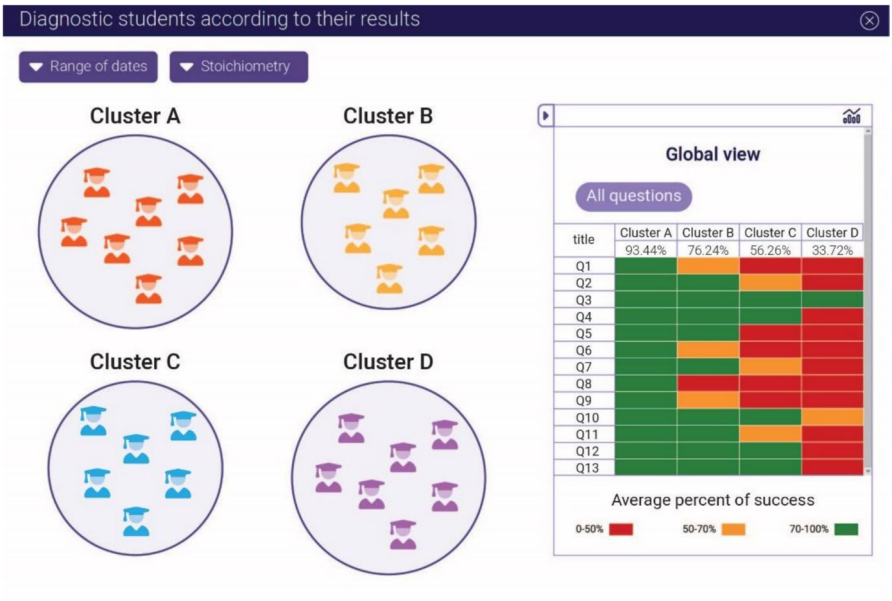

Fig. 3 Teacher dashboard

'struggling') but provide analytics on students' knowledge profiles with respect to a set of assessment items at a certain point in time. For instance, GrouPer might group students sharing a common misconception, encompassing a diverse range of knowledge, abilities, and skills within each cluster. The table in the dashboard provides more details about the response patterns that are mapped to each profile, with each column representing a specific cluster and each row indicating differences in student responses to particular items. This format enables teachers to distinguish between groups. The cells in the table are colored according to the following logic: green means that over 70% of the students in the cluster answered the item correctly, red indicates less than 50% correctness, and yellow represents intermediate cases. The color scheme was co-designed with teachers in a previous study (Nazaretsky et al., 2021). The procedures for fitting thresholds, deciding the number of clusters, etc., can be found in (Din et al., 2023; Nazaretsky et al., 2022a). The full protocol is presented in Appendix A.

It is important to note that throughout all stages of the research, the teachers had access to the assignment (see Appendix A), allowing them to potentially relate the identified clusters to the specific assignment if they thought it was necessary for them. Teachers were only introduced to the additional information provided by the two types of XAI after they had received this basic instruction regarding GrouPer. In the staged research protocol, the teachers received the explanations together with the system's recommendation. In the first stage, the teachers received the *data-driven* explanation. In the second stage, the *domain-driven* explanations were unveiled, and the teachers were asked again to assess their understandability, trust, and acceptance.

XAI Conditions: Data-driven and Domain-driven Explanations

The XAI conditions included two types of explanations presented to teachers gradually in order to evaluate their impact on understandability, trust, and acceptance. The first was a feature importance explanation, highlighting the most crucial questions for the model's results and providing a global explanation. The technical framework used to validate this explanation was developed in a previous study (Feldman-Maggor et al., 2024), where feature importance methods, which are typically applied to *supervised* machine learning, were adapted to generate a global explanation for *unsupervised* machine learning (cluster analysis).

The second was a semantic explanation describing in the curricular language that teachers' speak' the skills and competencies that each profile masters or struggles in, synthesized from the items that were most important for each cluster (see appendix A). These explanations, which we refer to as *domain-driven*, build upon the feature importance analysis. The rationale for focusing on these two levels of explanation was based on previous studies discussed in Sections "Theoretical Background" and "Research Hypothesis and Research Questions". The importance explanation acts as an information filter, highlighting the parts of the data that were most influential on the decision without providing synthesized meaning, and is thus referred to as '*data-driven.'* These types of explanations received the most attention in applications of XAI to AI-EdTech (e.g., Ghosh et al., 2024; Kar et al., 2023), but results on their interpretability by educators were mixed (Feldman-Maggor et al., 2024; Swamy et al., 2023). Studies from outside the education domain suggest that incorporating semantic domain information can significantly improve the quality of explanations (Donadello & Dragoni, 2020; Panigutti et al., 2020), thus referred to as '*domain-driven.'* We used this design to assess our research theory-driven hypothesis that domain-driven information can significantly enhance the quality of explanations compared to relying solely on data-driven ones (RQ2).

As explained in the previous section (The AI tool), GrouPer was used as a research vehicle. Validating its interface, which was co-designed with teachers (Nazaretsky et al., 2022a), was not part of the current research. Still, to minimize the possible influence of the specific interface on the participants' understanding, both types of explanations were also presented to the teachers in a textual format alongside GrouPer's dashboard (see Fig. 3 and Appendix A). The underlying rationale is based on previous HCI and XAI studies emphasizing supplementing visualizations with textual explanations (Haque et al., 2023).

Research Tools, Data Collection

Survey

After each stage of receiving an XAI explanation, teachers answered a short survey that included three items measuring perceived attributes: understandability, trust, and acceptance of the AI recommendations. The items were adapted from previous

studies that assess understandability, trust, and acceptance (Hadash et al., 2022; Nazaretsky et al., 2022b; Ribeiro et al., 2016). Each item was measured twice: the first measurement was after the data-driven explanation, and the second measurement was after adding the domain explanation (see Appendix A). The staged protocol and survey typically took about 45 min to complete, which is considered lengthy in real-world research contexts. To avoid participant fatigue and reduce attrition, the constructs of trust, acceptance, and understandability were evaluated quantitatively using a single item. We adopted this approach following previous study in the field of XAI (Ribeiro et al., 2016) and practices acceptable in the psychometric evaluation (Fisher et al., 2016; Gogol et al., 2014; Williamson & Kizilcec, 2021). The variables and the items are listed below:

Acceptance: was measured using the item "I would accept the groups offered by the "GrouPer" tool," a 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.
Understandability: was measured using the item "The explanation using the (data-driven/domain-driven) helped me understand the GrouPer tool more." 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.
Dynamic learned trust was measured using the item "The explanation using the (data-driven/domain-driven) helped me trust the GrouPer tool more." 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

Semi-Structured Think-Aloud Protocol

Think-aloud is a research method that grants insight into the cognitive processes underlying actions or decisions. It involves the immediate and spontaneous verbalization of thoughts while performing a task or solving a problem, as Ericsson and Simon (1998) described. In this method, participants are instructed to articulate everything that crosses their minds during the activity without interpreting or analyzing their thoughts. This approach ensures that participants express their thought processes in real-time. Additionally, it enables triangulating quantitative data, as Becker et al. (2022) noted. Different types of think-aloud methods vary in their level of prompting (Charters, 2003). The study employed a semi-structured think-aloud technique to allow follow-up questions. The semi-structured think-aloud protocol enables us to gather teachers' ideas about the explanations and their perceived attributes without explicitly steering their opinions with leading questions. At the same time, this approach allows probing participants'responses in more depth than written responses alone (Good et al., 2020). We follow Good et al. (2020), which suggested that, ideally, participants in a think-aloud study should not require coaching but should spontaneously verbalize their inner speech, as researcher modeling may introduce bias into think-aloud reporting. We follow Charters (2003), who emphasizes that researchers should allow participants' think-aloud behavior to remain as natural as possible, even if this results in varying degrees of information among participants. We chose this approach to

ensure that we could also collect quantitative data from these participants, as we will explain in the next section. This method was specifically chosen to explore teachers' perceived attributes about understandability, trust, and acceptance of an AI tool. It aimed to prompt honest opinions without leading the participants with suggestive questions while allowing for deeper probing than written responses permit, following the approach outlined by Good et al. (2020). The teachers who participated in this study were asked to describe their thoughts on each educational scenario and potential interfaces to an interviewer.

Data Analysis

In the quantitative analysis phase, we addressed RQ1 and RQ2. During the qualitative analysis phase, we provided additional support for the findings from RQ1 and RQ2 and also answered RQ3. This is shown in Fig. 4. Figure 4 represents part of Hoff and Bashir's (2015) model that was adopted in this research. In this study, we used "understandability" as a proxy for interpreting the results of XAI. As detailed in our literature review, understandability, in this context, serves as an operational stand-in for explainability. Hoff and Bashir (2015) note that design features significantly influence system performance, particularly regarding understandability. In our current study, we employed explainability as a key design feature, utilizing the concept of understandability to assess the system's performance. Consequently, the two types of XAI conditions—data-driven and domain-driven—refer to XAI design features that aim to achieve explainability. We incorporated these conditions into the model studied, as shown in Fig. 4.

Quantitative Analysis

We used a non-parametric test since the study variables are ordinal and not normally distributed (Villasenor Alva & Estrada, 2009). First, we applied the Mann–Whitney U test to determine if significant differences existed between the perceived attributes of the participants: those who completed the survey using the semi-think-aloud protocol and those who did not verbalize their thoughts. This analysis aimed to assess the possibility of aggregating the data and to check for any observer-expectancy effect. Second, we applied the Spearman correlation between the three ordinal variables (understandability, trust, and acceptance). We hypothesize a positive correlation exists between the three-variable building on Hoff and Bashir's (2013) and, therefore, use a one-tailed correlation test for each set of repeated measurements. We re-evaluated the significant correlations to show differences (α levels were adjusted by applying a Bonferroni corrected α-criterion of 0.025 (0.05/2). We interpret the strength of the relationship r as proposed by Xiao et al. (2016): r value that is above 0.5 is considered *high*, and r value in the range of 0.3–0.5 is considered *moderate.* We applied a one-tailed Wilcoxon signed-rank test to assess the changes between measurements aligning with our research hypothesis.

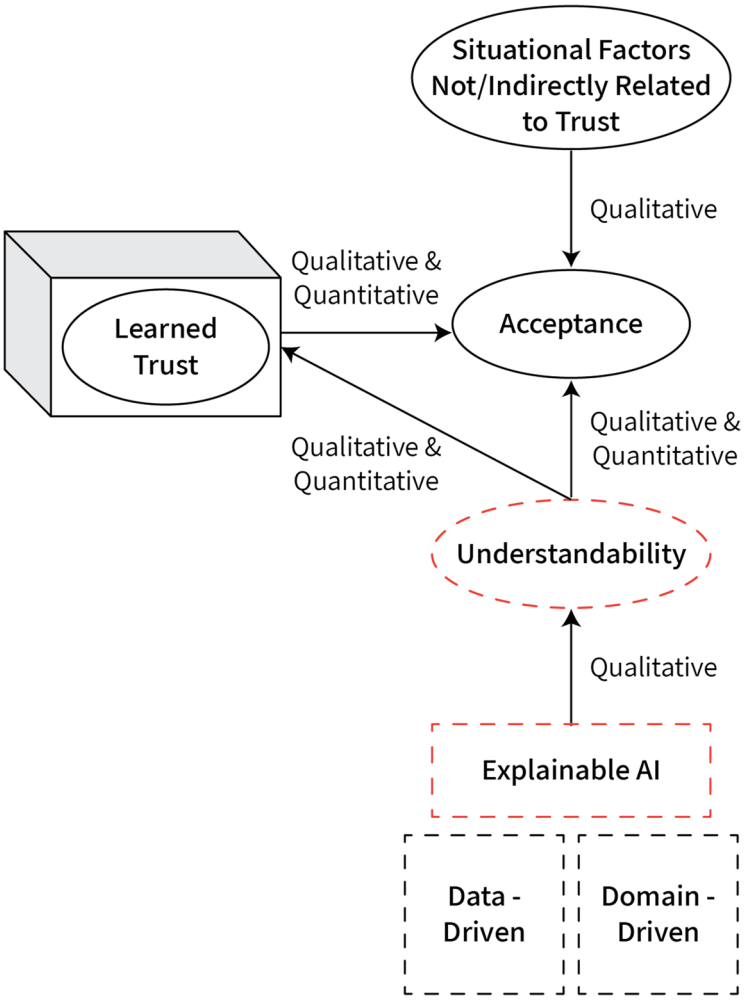

Fig. 4 Data analysis methods in relation to the appropriated conceptual model

Qualitative Analysis

The qualitative analysis was designed with two primary objectives: firstly, to complement and support the quantitative analysis conducted for answering RQ1 and RQ2, and secondly, to directly address and provide insights for RQ3. The qualitative analysis was conducted in line with the methodologies of Shkedi (2004) and Corbin and Strauss (1990), encompassing both First-order and Second-order theoretical analyses. In the First-order analysis, descriptive categories were created to capture the range of teachers' perceptions of the AI tool. This stage laid the foundation for the subsequent Second-order theoretical analysis, which built upon these initial analyses. The established descriptive categories were further analyzed and

interconnected, guided by trust, understandability, and acceptance (Cukurova et al., 2023; Hadash et al., 2022; Hoff & Bashir, 2013; Nazaretsky et al., 2022b). This approach enabled a deeper understanding of the underlying thematic structures in the context of AI's educational application. The validation process is aligned with Nowell et al. (2017) and Brod et al., 2009. The first author initially analyzed the data and identified various categories. The last author then validated this analysis through discussions, specifically to reach an agreement on whether particular perceptions aligned with trust, understandability, and acceptance. Following this, the first author re-analyzed the data, incorporating insights raised. The analysis was continuously refined and revalidated until both the first and last authors reached a consensus.

Results

The results section presents our findings in response to each research question. As outlined in the methodology, we conducted quantitative and qualitative analyses for RQ1 and RQ2 and a qualitative analysis for RQ3. The quantitative analysis is based on surveys completed by 41 teachers, with eleven of these participants also engaging in a qualitative think-aloud protocol. During data collection, we observed variations in the level of detail provided by these eleven teachers; some offered extensive insights, while others were more reserved. Unlike the interviews, where respondents answered specific questions, the think-aloud protocol allowed teachers to choose whether or not to address specific points (Charters, 2003). It is important to note that if a teacher does not explicitly provide their view on a particular issue, it does not necessarily mean they lack an opinion. Therefore, when we report the perspectives of 3 out of the 11 teachers who participated in the think-aloud sessions, it does not imply that the others disagree; they simply may not have articulated their stance. To ensure the validity of our findings, we focus on qualitative insights that were consistently observed among at least three teachers.

For the quantitative analysis, our objective was to collectively analyze all 41 teachers' responses. To ensure the validity of this approach, we conducted a preliminary Mann–Whitney U test to determine if there were significant differences in perceived attributes towards trust, understandability, and acceptance between teachers who participated in the think-aloud protocol and those who did not. The lack of significant differences allowed us to aggregate the data, facilitating a comprehensive quantitative analysis of all the responses.

RQ1: What is the Relationship Between Understandability, Trust, and Acceptance, and do These Variables Change During the Teacher's Interaction with the System?

We begin by addressing RQ1, which assesses the relationship between teachers' understandability, trust, and acceptance and examines whether these factors changed after their interactions with the system.

A one-tailed Spearman's correlation analysis revealed a highly positive and statistically significant correlation between trust, understandability, and acceptance.

The results are presented in Fig. 5A and B. Figure 5A illustrates the correlations between the variables following the presentation of the data-driven explanation, while Fig. 5B shows the correlations after the domain-driven explanation was presented. As we can see in Fig. 5A and B, all correlations were high, above 0.5, except one moderate correlation with a value of 0.439. The strong positive correlations indicate a significant relationship between understandability, trust, and acceptance. Notably, Fig. 5A and B demonstrate that the correlations between the variables were stronger following the Domain-driven explanation, suggesting that these factors vary during the teachers' interactions with the system. We further explore these changes in the context of answering RQ2.

We also gained insights into the relationship between understandability, trust, and acceptance through our qualitative analysis, as described below. Specifically, we examined the relationship between 1) understandability and trust, 2) understandability and acceptance, and 3) trust and acceptance. Below, we present how analyzing the responses from the think-aloud protocol supports the quantitative findings.

As a preliminary step, we wanted to first establish the contribution of the XAI features to teachers' perceived understanding of the analysis, namely, to examine the functionality of the explanations. Analyzing the protocol yielded that seven teachers reported that the XAI features increased their understandability of the AI tool recommendations. As an example, teacher #1 noted that "*Information about the three questions that had the most impact helped me to understand the division of groups better*". The three questions represent the feature importance explanation, which serves as a data-driven explanation. Since the teacher mentioned that this explanation helped her understand the division between the groups, we consider this an example that provides evidence that XAI can enhance perceived understandability. Next, we sought evidence that understandability enables teachers to validate the results. We focus on this issue because, in Hoff and Bashir's model (2013, 2015), the ability to validate the system performance establishes trust, and this ability relates to understandability. From five teachers, we learned that understandability enabled them to validate the AI group's recommendations through statistics or domain expertise. For example, one teacher (#4) noted, "*At*

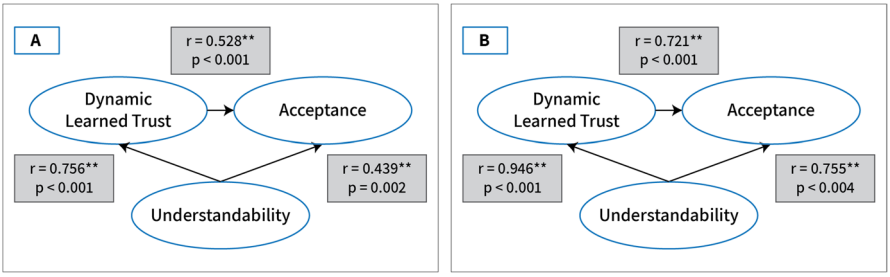

Fig. 5 Correlation between variables. **A** – XAI = Data-driven, **B** – XAI = Domain-driven (N = 41). ** Comparison is significant at the 0.01 level (one-tailed). A Bonferroni correction was applied to the α value to determine significance

least I'll look at samples as they do in factories. The machine does everything; I am just taking samples and checking if the tool did it the same way that I did".

Overall, the qualitative analysis supported our working assumption that the XAI features are effective in increasing understandability and that teachers are relying on understandability, as Hoff and Bashir's model (2013, 2015) proposes, to validate the system performance. Furthermore, we assessed indications of the relationship between understandability and the development of trust among teachers, observing whether their perceived trust changed following their reports of perceived understandability. We examined teachers' responses as indicators of changes in their perceived attributes toward trust. Seven out of eleven teachers who participated in the think-aloud protocol expressed that the explanation provided improved their trust in the AI tool. However, this analysis also revealed that the relation between understandability and trust is not straightforward: For three teachers, understandability alone was insufficient to establish trust in the AI tool. These teachers reported that, for them, real classroom experience is needed to gain trust in the system. As one of these teachers (#5) said: *"The more I will experiment and obtain results that align with my expectations, the more I will trust the system."*

Next, we evaluated the relationship between understandability and acceptance. We observed explicit expressions of the relationship between understandability and acceptance among four teachers. For instance, one teacher (10#) said, *"So, what do I do with this information? OK, knowing that this question is difficult for one of the groups informs me about what I need to concentrate and work on (with the students in each group, indicating that the teacher is accepting the AI recommendation)"*.

Finally, we looked for indications of the relationship between trust and acceptance. This indication emerged from three teachers. One teacher (#9) stated, *"I generally trust these tools. I accept a certain margin of error, as no tool is perfect. Even if the tool (GrouPer) has only 20% reliability, it's still valuable. It saves time by helping target students based on specific skills or content areas in chemistry. I trust the tool (GrouPer), knowing its effectiveness will improve as more data becomes available"*. This teacher explained the logic behind her trust, stating that she accepts its analysis, although it may not be perfect, as she has positive expectations that it will improve her teaching.

RQ2: To What Extent do Teachers' Perceptions of System Performance, Trust, and Acceptance of AI Tools Shift after Receiving Domain-driven Explanations Following Data-driven Explanations?

In this section, we explore RQ2 by examining how different types of XAI, specifically receiving domain-driven explanations following data-driven explanations, shift teachers'understandability of system performance, trust, and acceptance of AI tools.

The changes in teachers' perceived attributes during their interactions with the AI system are presented in Figs. 6 and 7. These figures present an increase in understandability, trust, and acceptance of AI when transitioning from explanations that are based solely on data-driven methods to explanations that are domain-driven. This increase in teachers' perceived attributes towards XAI, an increase from relying on

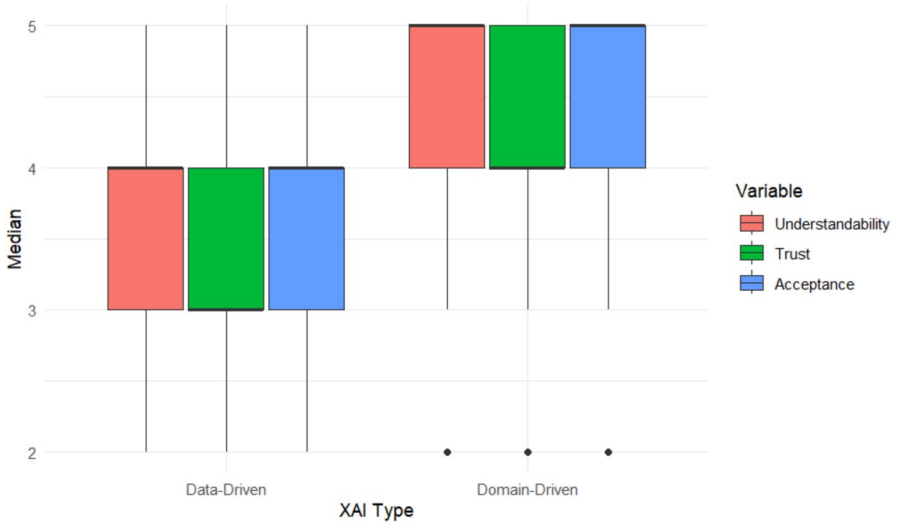

Fig. 6 Differences in the three perceived attributes according to XAI type, as shown by a boxplot

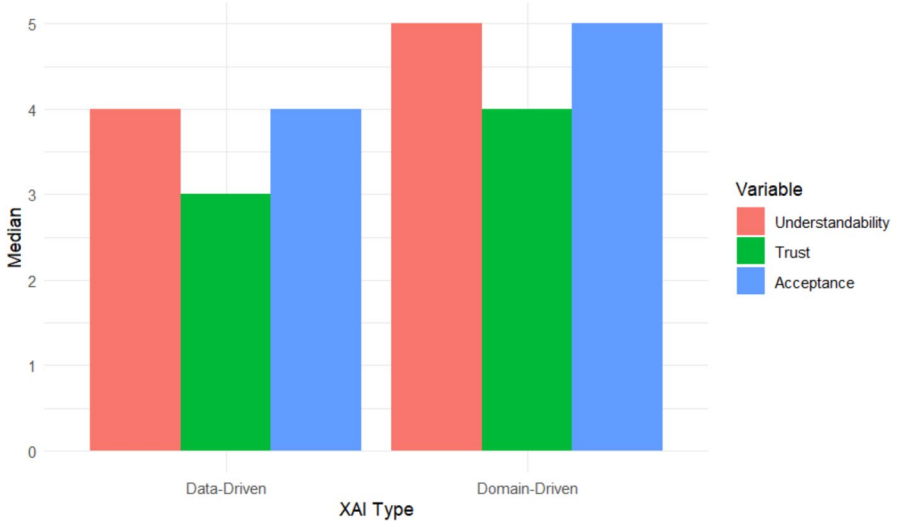

Fig. 7 Differences in the three perceived attributes according to XAI type, as shown by a barplot

feature importance to integrating domain-driven, was statistically significant. Specifically, our research hypothesis was that leveraging domain-driven explanations will improve the quality of explanations compared to relying solely on data-driven explanations, increasing understandability, trust, and Acceptance. To test this, we employed a one-tailed Wilcoxon signed-rank test to compare the values measured for these constructs after the data-driven explanations and after domain-driven ones.

The test yielded statistically significant results for all the variables: Understandability—W = 80.5, $p = 0.005$**; Learned trust—W = 52, $p = 0.002$**; Acceptance—W = 22.5, $p = 0.003$**.

We next turned to qualitative analysis to triangulate and better understand the quantitative results. As we discussed in our response to RQ1, seven teachers indicated that XAI enhances their understandability of the AI tool's recommendations. Specifically, among these seven, three teachers noted that data-driven explanations increased their trust in the AI tool. However, all seven teachers found that domain-driven explanations were more influential in building their trust in the tool. This suggests that some teachers prefer domain-driven explanations over data-driven ones in fostering trust. For example, one teacher (#3) commented, "*Semantic information about the difficulties and strengths helped me trust more*". Similarly, another teacher (#4), who initially expressed skepticism (after receiving the data-driven explanations), acknowledged a shift in her acceptance after receiving the domain-driven explanations: "*At first, I was doubtful, but now I am convinced of the tool's utility, especially in mapping students*".

RQ3: Which Additional Factors May Influence Teachers' Willingness to Accept and Rely on AI Analysis?

We addressed RQ3 qualitatively by evaluating which additional factors may influence teachers' willingness to accept AI recommendations. We identified two categories highlighted by teachers as factors promoting acceptance. We interpreted them as situational factors that influenced acceptance and were not directly related to trust. These two categories are pedagogical considerations and AI's workload reduction potential. The category of pedagogical considerations emerged in eight think-aloud protocols. For example, as noted by one teacher (#1), *"In this approach, the learning material is broken down to allow some students to progress at a pace, that's right for them, even if it's slower than the rest of the class. This method supports a personal learning pace, ensuring that all students can advance in a way that suits their individual needs"*. The role of AI as workload reduction potential' was described by six teachers. For instance, a teacher (#5) explains: "*Mapping students based on specific criteria, such as their proficiency in chemistry, can be time-consuming for me… It helps identify students needing more attention or support in certain areas"*.

The qualitative analysis presented in the results section represents our second-order data analysis. These results have been incorporated into our theoretical framework, as shown in Fig. 8, where we added the option to validate AI recommendations by experience or understandability of the original model. In our theoretical framework, trust and understandability impact acceptance. However, we acknowledge that acceptance is also influenced by situational factors that are not directly related to trust. Our qualitative analysis identified potential workload reduction and pedagogical considerations as additional factors unrelated to trust that may impact acceptance. For example, workload reduction is likely to enhance acceptance, and pedagogical considerations can lead to enhanced acceptance if teachers perceive the

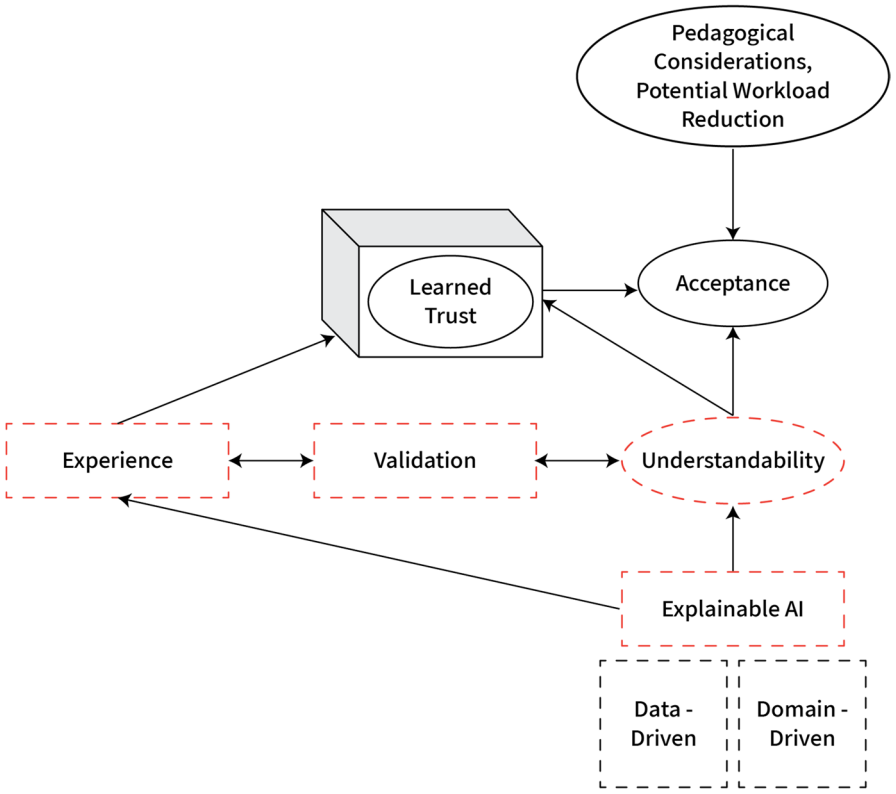

Fig. 8 Conceptual framework and second-order data analysis

system as aligned with their pedagogical approach. However, these two factors are certainly only a subset of the additional situational factors that impact acceptance.

Discussion

Trust is key to the teachers' acceptance of AI-EdTech recommendations. Research in other domains underlined the dynamic nature of trust, and studying this aspect within AIED contexts with teachers was the first goal of this research. Specifically, previous AIED research on trust (e.g., Conijn et al., 2023; Nazaretsky et al., 2022a) assumed that increasing knowledge and transparency by explaining how the AI system reaches its decisions will increase trust and acceptance. We concur with the approach, but our research sought to provide a more structured, theory-driven explanation highlighting the dynamic nature of trust with its different categories of dispositional, situational, and learned trust. We considered the model of Hoff and Bashir (2013, 2015; see Fig. 1) on trust in automation particularly appropriate to our purposes. The original model describes the relationship between design features,

system performance, learned trust, and acceptance. We adapt it to study the relations between XAI, understandability of the system performance, trust in it, and acceptance of its decisions in the following manner. First, we observe that understandability is an aspect of system performance (which Hoff & Bashir, 2013, 2015 also mentioned explicitly). We then make the additional observation that *XAI features are Design Features meant to increase understandability*. These two refinements of the model are illustrated in Fig. 2. This leads to the refined model presented in Fig. 4, which relates to XAI, understandability, trust, and acceptance (hereafter, we refer to this refined model as '*the model*').

Our first research question then concentrated on validating the model's connections between understandability, trust, and acceptance. We did so in empirical research with teachers to evaluate the applicability of the model to explain important dimensions in teacher-AI interactions. By doing so in varied conditions, we hoped also to shed light on the dynamic nature of trust. The quantitative analysis of RQ1 (see Fig. 5) reveals i) a strong connection between understandability, trust, and acceptance, as anticipated by the model, and ii) shows how trust changes as a function of temporal system-related aspects.

Next, in RQ2, we explored the relationship between XAI and understandability. To recap, our model interprets XAI as a system design feature that, according to the model, influences understandability, which impacts trust. This theoretical model can explain the findings of Nazaretsky et al., (2022a, 2022b, 2022c), who showed that understanding the internal logic behind how an AI grading system scores student responses increased teacher trust in the automated assessment. However, following the more nuanced results of Kizilcec (2016), who provided a perspective on the possibly opposite effect of different types of explanations on trust, we decided to examine the relation between XAI and understandability under several conditions, namely, with different types of XAI schemes.

The specific example of understandability in our study concerned the clustering results. Since clustering is based on an unsupervised algorithm, the results are not pre-defined and can be difficult to explain (Moshkovitz et al., 2020). This algorithmic approach to grouping differs from teachers'usual classroom practice, where groups are typically formed based on a range of grades (Boaler, 2020; Betts & Shkolnik, 2000). The findings from the think-aloud protocol provided evidence that teachers who received the data-driven explanation understood that the groups were not pre-defined but rather determined by the algorithm. Additionally, the domain-driven explanation helped them relate the clustering results to their existing knowledge and pedagogical practices. From this perspective, a combination of both explanations may be necessary for better understandability and trust formation.

The analysis of RQ1 and RQ2 was confirmatory and primarily quantitative, but it was also triangulated with the think-aloud protocols, which provided richer data that included additional insights. When analyzing the think-aloud transcripts, it was evident that additional factors, which were not included in the model, might influence trust. Addressing this aspect was formulated through RQ3. To answer this, we conducted an exploratory, qualitative analysis to reveal additional factors that may play a role in shaping trust and acceptance. This analysis highlighted additional factors, potential workload reduction and,

pedagogical considerations as situational factors. These factors were also highlighted in previous research as significant factors influencing the adoption of AI-EdTech with a survey approach (Cukurova et al., 2023), yet our work adds more nuance to these factors with its qualitative nature.

In terms of the strength of the influence that each factor has on trust, the current study did not try to evaluate the relative influence XAI has on trust versus other factors. However, a few teachers stressed their need for a real experience with their own classroom using GrouPer. These teachers reported that real classroom experience would be necessary to develop trust (see Fig. 8). Furthermore, our analysis provides evidence that understandability achieved through XAI is highly correlated with acceptance of AI-EdTech recommendations. This is in line with the work of Hancock et al. (2011), who argued that system features are more important than individual features in determining the amount of trust one would have in the system. Thus, our work provides strong evidence that justifies the investment in XAI as a means to increase teachers' trust in and acceptance of AI-EdTech recommendations.

Research Limitations

The research has several limitations. One threat to the internal validity was the within-subject staged design that was used to evaluate the level of and the relation between understandability, trust, and acceptance under the two XAI conditions, domain-driven and data-driven explanations. The main risk with this type of design is a 'carryover effect' in which the second stage of the research is impacted by the first one. We chose a within-subject design since our aim was to evaluate the adapted Hoff and Bashir (2013, 2015) model in the AIED context, which centers on how attitudes change over time. A typical solution is conducting the research with multiple groups that receive the intervention in different orders. However, this solution requires a significantly larger number of research subjects, which was not feasible, as the 'authenticity' criterion imposed requirements that limited the number of suitable research subjects. In terms of the construct validity of our measurement, our model refers to understandability, which is an attribute of the system, but our measurement of this attribute is through its perception by the participants, which may differ from its actual one. Also, with respect to the measurement, the variables of trust, acceptance, and understandability were evaluated using a single item. This approach was chosen to avoid overloading the participants, as the staged protocol and survey typically took about 45 min to complete, which is considered long in real-world research contexts. We note that it was important to conduct this research in a real-life context, as the authenticity of the setting was found to influence teachers'attitudes, thereby impacting the validity of the results (Nazaretsky et al., 2022a). Thus, using a single item was a deliberate choice aimed at balancing between two competing threats to the validity of the research. Lastly, we acknowledge that the small sample size poses limitations to the generalizability of the results.

Implications and Future Research

This study has several practical applications. With respect to XAI design, developers should bear in mind that the understandability of XAI features influences trust and acceptance in AI and that effective explanations should include technical, data-driven insights and domain-driven ones that educators can interpret. Not surprisingly, the findings also emphasize that teachers are more likely to trust and accept AI tools if they align with their pedagogical practices and help reduce their workload. Regarding teachers' professional development, while this was not the focus of our research, we believe one important implication is that AI-related training should also enhance educators'data fluency, enabling them to make better use of more technical, data-driven explanations.

In addition to practical implications, several avenues for future research can be sketched. In this study, we studied trust through the prism of the model proposed by Hoff and Bashir (2013, 2015), which centers on the dynamic nature of trust and its relation to XAI and understandability. However, trust and its dynamics can also be impacted by other factors. For example, trust can be learned through repeated experience with a certain tool (e.g., learning to trust a navigation system after being repeatedly convinced of its accuracy) or through the credibility of the institution behind the tool (e.g., trusting a drug, knowing that the FDA approved it). Future research can further investigate these factors.

Another direction for future research is the dynamic nature of other types of attitudes/perceived attributes, beyond trust, toward AI tools in education. While Hoff and Bashir (2013, 2015) explicitly used the term "dynamic" only in the context of learned trust, our findings suggest that understandability and acceptance, both of which are associated with trust, can also evolve during and after the interaction with the AI system. Future research could further investigate the dynamic nature of these and other AI-related attitudes in the context of education.

Our study focused on the type of information provided through the explanations rather than on how the information was served to the educators. Future research could also explore how different dimensions beyond the *type of information*, such as types of visualizations, impact educational stakeholders'understandability. Overall, while the understandability of AI recommendation systems has been studied in healthcare, transportation, and HCI (Joyce et al., 2023; Perez-Cerrolaza et al., 2024; Shin, 2021), it has received less attention within the AIED community. Since understandability is closely linked to trust, understanding can help ensure that trust in AI is developed through informed understanding rather than based solely on intuitive impressions or positive user experiences. Finally, while we have focused on teachers, future research can study trust among other educational stakeholders, such as students, policymakers, and parents. This is particularly relevant today, as policymakers actively shape strategies for AI implementation in education (Ifenthaler et al., 2024), and the public gains widespread access to various generative AI tools, influencing both policy discourse and classroom practices.

Summary and Contribution

Trust is a critical factor in teachers' adoption of AI-EdTech, yet its dynamic nature remains underexplored, and there is a lack of well-established theory connecting it to XAI. To address this gap, the present study first proposed a theoretical model that connects XAI, understandability, trust, and acceptance and is based on applying Hoff and Bashir's Trust in Automation model (2013, 2015) to the space of teacher-AI interaction. Then, the study explored the relationship between different types of XAI schemes and teachers'trust and acceptance of AI-powered recommendations. According to the model, these factors are linked to the system's understandability, which is achieved through the XAI schemes. The model was tested through a mixed-method, within-subject study involving 41 in-service chemistry teachers who performed an authentic task with an AI-powered EdTech tool. Participants were exposed to two types of XAI explanations: data-driven (feature importance) and domain-driven (semantic explanations in curricular language). The results confirmed a strong positive correlation between understandability, trust, and acceptance, as predicted by the model. Teachers demonstrated a preference for domain-driven explanations, which significantly enhanced their perceived understandability, leading to a positive change in trust and acceptance. The first key contribution of the present research is thus proposing a theoretical explanation of how XAI can contribute to trust and acceptance by increasing the system's understandability. The second key contribution is showing that domain-driven explanations can further enhance teachers' understandability, trust, and acceptance compared to explanations that are merely data-driven. In doing so, the findings stress the dynamic nature of teachers' trust in AI systems, showing that it evolves with user experience and the suitability of the XAI scheme, which is the third research contribution. Furthermore, two contextual factors, pedagogical perspectives and workload reduction potential, emerged as additional factors that may influence teachers' acceptance of AI recommendations.

Appendix A. The Educational Protocol Provided to the Teachers

Introduction to the Tool

Scenario: Students completed a questionnaire administered via an online environment. The responses were processed using a clustering algorithm, which divided the class into groups (clusters) of students with similar characteristics. The results were presented in the GrouPer dashboard.

Goals of the GrouPer Tool:

- Provide immediate diagnosis of students' strengths and difficulties across various aspects.
- Automatically group students based on common characteristics.

We will now present a scenario demonstrating how the GrouPer tool works. In this scenario, 11 th-grade students completed a matriculation preparation questionnaire on the topic of stoichiometry (click here for the questions.).

The 11 th-grade students' responses were analyzed, and the dashboard displays the grouping (clustering) results. Each group (cluster) contains students with similar response patterns. The dashboard shows the number of students in each group. The GrouPer Dashboard is presented in Fig. 9.

GrouPer is an AI-powered tool utilizing large-scale data, applying a machine learning algorithm to divide students into groups based on the similarity in their response patterns in a specific assignment. After analyzing many student responses, the algorithm forms groups and assigns each student to their most appropriate group. The circles in the dashboard (Fig. 9) represent students from the teacher's own class. It is important to note that the clusters presented do not reflect the students' general abilities (like 'excellent' or 'struggling') but provide analytics on students' knowledge profiles concerning a set of assessment items at a certain time.

The table, which uses green, orange, and red colors, displays the representative response pattern for each profile. Each column represents a specific cluster (group). Each row corresponds to a particular questionnaire item, showing how students in each cluster responded to that item. The color coding works as follows: Green indicates that more than 70% of students in the cluster answered the item correctly. Red indicates that fewer than 50% answered correctly. Orange represents cases in between (50%-70% correct).

Explanations about artificial intelligence algorithms are usually designed for experts in the field. However, as AI-based tools become more common in

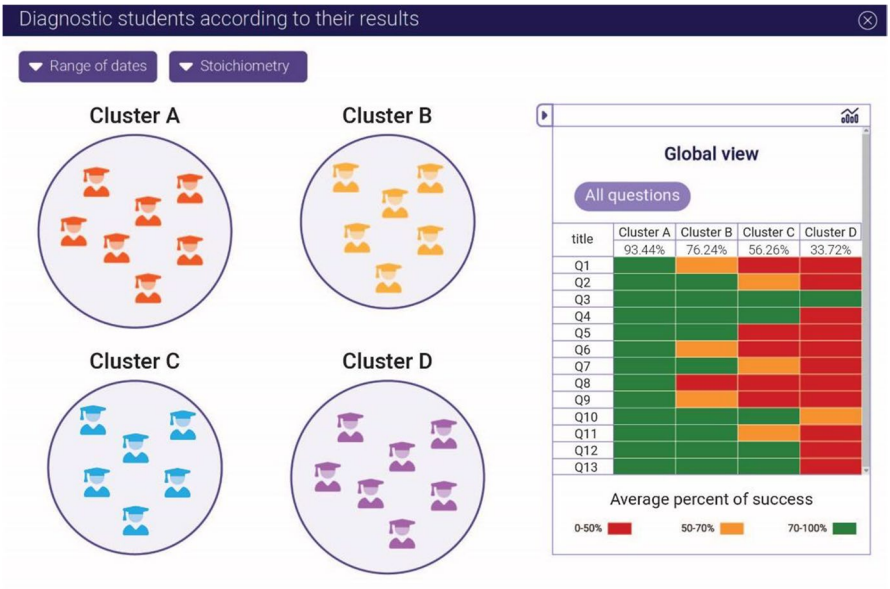

Fig. 9 GrouPer Dashboard

educational settings, there is a growing need to make these explanations accessible to teachers and other users. One widely accepted approach for data-driven explaining how an AI-based algorithm works is to highlight which questions had the greatest influence on the model's clustering decisions. In the case of the stoichiometry questionnaire, questions 8, 9, and 11 had the most significant impact on determining a student's group. (This analysis was conducted using several different methods.).

For each of the following statements, please indicate the extent to which you agree using a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

I would accept the groups offered by the "GrouPer" tool, a 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

The explanation using the data-driven helped me understand the GrouPer tool more. 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

The explanation using the data-driven helped me trust the GrouPer tool more. 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

In order to improve the GrouPer tool, the following domain-driven explanations were added, which include additional information about the clusters. (See Fig. 10).

For each of the following statements, please indicate the extent to which you agree using a Likert scale from 1 (strongly disagree) to 5 (strongly agree).

I would accept the groups offered by the "GrouPer" tool, a 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

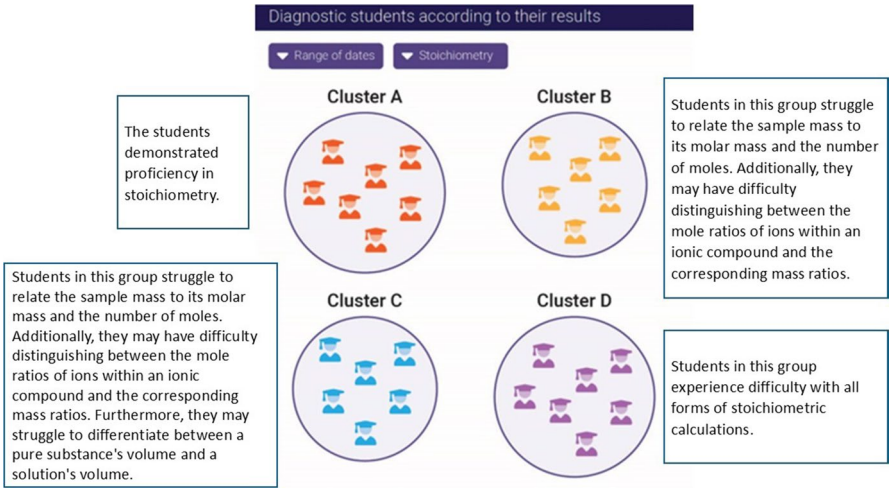

Fig. 10 Domain-driven explanation for each cluster

The explanation using the domain-driven helped me understand the GrouPer tool more. 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

The explanation using the domain-driven helped me trust the GrouPer tool more. 5-point scale from 1 = Strongly disagree, 2 = Disagree, 3 = Neutral, 4 = Agree, and 5 = Strongly agree.

Author Contribution All authors, Yael Feldman-Maggor, Mutlu Cukurova, Carmel Kent, and Giora Alexandron, conceptualized and designed the study. Yael Feldman-Maggor wrote the original draft, and Giora Alexandron supervised the study. All authors edited the manuscript.

Funding Open access funding provided by Weizmann Institute of Science. The authors declare they have no financial interests.

Data Availability Data availability upon request.

Declarations

Conflict of interest The authors declare no competing interests.

References

Antonietti, C., Cattaneo, A., & Amenduni, F. (2022). Can teachers' digital competence influence technology acceptance in vocational education? *Computers in Human Behavior, 132*, 107266. https://doi.org/10.1016/j.chb.2022.107266

Becker, S., Obersteiner, A., & Dreher, A. (2022). Eye tracking–promising method for analyzing mathematics teachers assessment competencies? *In 2022 Symposium on Eye Tracking Research and Applications* (pp. 1–4). https://doi.org/10.1145/3517031.3529244

Betts, J. R., & Shkolnik, J. L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review, 19*(1), 21–26. https://doi.org/10.1016/S0272-7757(99)00022-9

Boaler, J. (2020). Ability Grouping in Mathematics Classrooms. In Lerman, S. (ed) *Encyclopedia of Mathematics Education*. Springer. https://doi.org/10.1007/978-3-030-15789-0_145

Brod, M., Tesler, L. E., & Christensen, T. L. (2009). Qualitative research and content validity: Developing best practices based on science and experience. *Quality of Life Research, 18*, 1263–1278. https://doi.org/10.1007/s11136-009-9540-9

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *In 2015 international conference on healthcare informatics* (pp. 160–169). IEEE. https://doi.org/10.1109/ICHI.2015.26

Celik, I. (2023). Towards Intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior, 138*, 107468. https://doi.org/10.1016/j.chb.2022.107468

Charness, G., Gneezy, U., & Kuhn, M. A. (2012). Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization, 81*(1), 1–8. https://doi.org/10.1016/j.jebo.2011.08.009

Charters, E. (2003). The use of think-aloud methods in qualitative research an introduction to think-aloud methods. Brock Education Journal, 12(2). was added to the references list. https://doi.org/10.26522/brocked.v12i2.38

Chaudhry, M.A., Cukurova, M., Luckin, R. (2022). A Transparency Index Framework for AI in Education. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, spsampsps V. Dimitrova, V. (eds) Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium. AIED 2022. Lecture Notes in Computer Science, vol 13356. Springer, Cham. https://doi.org/10.1007/978-3-031-11647-6_33

Choi, S., Jang, Y., & Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational artificial intelligence tools. *International Journal of Human-Computer Interaction, 39*(4), 910–922. https://doi.org/10.1080/10447318.2022.2049145

Choung, H., David, P., & Ross, A. (2023). Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human-Computer Interaction, 39*(9), 1727–1739. https://doi.org/10.1080/10447318.2022.2050543

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics, 10*(1), 37–53. https://doi.org/10.18608/jla.2023.7801

Corbin, J. M., & Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative Sociology, 13*(1), 3–21.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies, 58*(6), 737–758. https://doi.org/10.1016/S1071-5819(03)00041-7

Cukurova, M., Luckin, R., & Kent, C. (2020). Impact of an artificial intelligence research frame on the perceived credibility of educational research evidence. *International Journal of Artificial Intelligence in Education, 30*, 205–235. https://doi.org/10.1007/s40593-019-00188-w

Cukurova, M., Miao, X., Brooker, R. (2023). Adoption of Artificial Intelligence in Schools: Unveiling Factors Influencing Teachers' Engagement. In N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, spsampsps V. Dimitrova (Eds.) Artificial Intelligence in Education. AIED 2023. Lecture Notes in Computer Science (pp. 151–163), vol 13916. Springer, Cham. https://doi.org/10.1007/978-3-031-36272-9_13

Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *International Journal of Human-Computer Studies, 162*, 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

Din, B., Nazaretsky, T., Feldman-Maggor, Y., Alexandron, G. (2023). Automated identification and validation of the optimal number of knowledge profiles in student response data. In M. Feng, T. K¨aser, & P. Talukdar (Eds.), *Proceedings of the 16th International Conference on Educational Data Mining* (pp 458–465). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.8115744

Donadello, I., spsampsps Dragoni, M. (2020, November). SeXAI: a semantic explainable artificial intelligence framework. In International Conference of the Italian Association for Artificial Intelligence (pp. 51–66). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-77091-4_4

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Feldman-Maggor, Y., Nazaretsky, T., & Alexandron, G. (2024). Explainable AI for unsupervised machine learning: A proposed scheme applied to a case study with science teachers. In Proceedings of 16th international conferences of computer supported education (CSEDU). https://doi.org/10.5220/0012687000003693

Fisher, G. G., Matthews, R. A., & Gibbons, A. M. (2016). Developing and investigating the use of single-item measures in organizational research. *Journal of Occupational Health Psychology, 21*(1), 3–23. https://doi.org/10.1037/a0039139

Fréchette, A., Kotthoff, L., Michalak, T., Rahwan, T., Hoos, H., & Leyton-Brown, K. (2016). Using the shapley value to analyze algorithm portfolios. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1). https://doi.org/10.1609/aaai.v30i1.10440

Ghosh, S., Kamal, M. S., Chowdhury, L., Neogi, B., Dey, N., & Sherratt, R. S. (2024). Explainable AI to understand study interest of engineering students. *Education and Information Technologies, 29*, 4657–4672. https://doi.org/10.1007/s10639-023-11943-x

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80–89). IEEE. https://doi.org/10.1109/DSAA.2018.00018

Gogol, K., Brunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischbach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology, 39*(3), 188–205. https://doi.org/10.1016/j.cedpsych.2014.04.002

Good, M., Marshman, E., Yerushalmi, E., & Singh, C. (2020). Graduate teaching "assistants' views of broken-into-parts physics problems: Preference for guidance overshadows development of self-reliance in problem solving. *Physical Review Physics Education Research, 16*(1), 010128. https://doi.org/10.1103/PhysRevPhysEducRes.16.010128

Guleria, P., & Sood, M. (2023). Explainable AI and machine learning: Performance evaluation and explainability of classifiers on educational data mining inspired career counseling. *Education and Information Technologies, 28*(1), 1081–1116. https://doi.org/10.1007/s10639-022-11221-2

Hadash, S., Willemsen, M. C., Snijders, C., & IJsselsteijn, W. A. (2022, April). Improving understandability of feature contributions in model-agnostic explainable AI tools. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (pp. 1–9). https://doi.org/10.1145/3491102.3517650

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Haque, A. B., Islam, A. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change, 186*, 122120.

Hoff, K., & Bashir, M. (2013). A theoretical model for trust in automated systems. In CHI'13 extended abstracts on human factors in computing systems (pp. 115–120). https://doi.org/10.1145/2468356.2468378

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Huang, H. Y., & Bashir, M. (2017). Users' adoption of mental health apps: Examining the impact of information cues. *JMIR mHealth and uHealth, 5*(6), e6827. https://doi.org/10.2196/mhealth.6827

Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*, 1–18. https://doi.org/10.1007/s10758-024-09747-0

Jang, Y., Choi, S., Jung, H., & Kim, H. (2022). Practical early prediction of students' performance using machine learning and eXplainable AI. *Education and Information Technologies, 27*(9), 12855–12889. https://doi.org/10.1007/s10639-022-11120-6

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Johnson, D. G., & Verdicchio, M. (2017). AI anxiety. *Journal of the Association for Information Science and Technology, 68*(9), 2267–2270. https://doi.org/10.1002/asi.23867

Joyce, D. W., Kormilitzin, A., Smith, K. A., & Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *Npj Digital Medicine, 6*(1), 6. https://doi.org/10.1038/s41746-023-00751-9

Kar, S. P., Das, A. K., Chatterjee, R., & Mandal, J. K. (2023). Assessment of learning parameters for students' adaptability in online education using machine learning and explainable AI. Education and Information Technologies, 1–16. https://doi.org/10.1007/s10639-023-12111-x

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., Knight, S. B., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence, 3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2390–2395). https://doi.org/10.1145/2858036.2858402

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Leech, N. L., & Onwuegbuzie, A. J. (2007). An array of qualitative data analysis tools: A call for data analysis triangulation. *School Psychology Quarterly, 22*(4), 557. https://doi.org/10.1037/1045-3830.22.4.557

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: informing design practices for explainable AI user experiences. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1–15). https://doi.org/10.1145/3313831.3376590

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology, 14*(1), 1–59. https://doi.org/10.1145/3546872

Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. *Electronic Markets, 32*(4), 1993–2020. https://doi.org/10.1007/s12525-022-00605-4

Lumineau, F., & Schilke, O. (2020). Trust development across levels of analysis: An embedded-agency perspective. In Multilevel Trust in Organizations (pp. 102–112). Routledge. https://doi.org/10.4324/9781003029526-6

Maheshwari, G. (2023). Factors influencing students' intention to adopt and use ChatGPT in higher education: A study in the Vietnamese context. Education and Information Technologies, 1–29. https://doi.org/10.1007/s10639-023-12333-z

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I "don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human Factors, 55*(3), 520–534. https://doi.org/10.1177/0018720812465081

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS), 11*(3–4), 1–45. https://doi.org/10.1145/3387166

Moshkovitz, M., Dasgupta, S., Rashtchian, C., & Frost, N. (2020). Explainable k-means and k-medians clustering. *Proceedings of the 37th International Conference on Machine Learning*, Online, PMLR 119, 2020.

Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation. arXiv preprint arXiv:1802.00682.

Nazaretsky, T., Hershkovitz, S., & Alexandron, G. (2019). *Kappa Learning: A New Item-Similarity Method for Clustering Educational Items from Response Data*. International Educational Data Mining Society.

Nazaretsky, T., Cukurova, M., Ariely, M., & Alexandron, G. (2021). Confirmation bias and trust: human factors that influence teachers' attitudes towards AI-based educational technology. In *AI for Blended-Learning: Empowering teachers in real classrooms co-located with 16th European Conference on Technology Enhanced Learning (ECTEL 2021)*.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022a). Teachers trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology, 53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022b). An instrument for measuring teachers' trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 56–66). https://doi.org/10.1145/3506860.3506866

Nazaretsky, T., Bar, C., Walter, M., & Alexandron, G. (2022c). Empowering teachers with AI: Co-designing a learning analytics tool for personalized instruction in the science classroom. In *LAK22: 12th International Learning Analytics and Knowledge Conference* (pp. 1–12). https://doi.org/10.1145/3506860.3506861

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence*, 100368. https://doi.org/10.1016/j.caeai.2025.100368

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods, 16*(1), 1609406917733847.

OECD. (2021). *Tools for trustworthy AI: A framework to compare implementation tools for trustworthy AI systems*. OECD Digital Economy Papers, 312. Retrieved April 25th, 2025, from: https://doi.org/10.1787/008232ec-en

Panigutti, C., Perotti, A., & Pedreschi, D. (2020, January). Doctor X.A.I.: an ontology-based approach to black-box sequential data classification explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 629–639). https://doi.org/10.1145/3351095.3372855

Perez-Cerrolaza, J., Abella, J., Borg, M., Donzella, C., Cerquides, J., Cazorla, F. J., Englund, C., Tauber, M., Nikolakopoulos, G., & Flores, J. L. (2024). Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey. *ACM Computing Surveys, 56*(7), 1–40.

Qin, F., Li, K., & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology, 51*(5), 1693–1710. https://doi.org/10.1111/bjet.12994

Quinn, T. P., Senadeera, M., Jacobs, S., Coghlan, S., & Le, V. (2021). Trust and medical AI: The challenges we face and the expertise needed to overcome them. *Journal of the American Medical Informatics Association, 28*(4), 890–894. https://doi.org/10.1093/jamia/ocaa268

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should i trust you"?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Saif, N., Khan, S. U., Shaheen, I., ALotaibi, A., Alnfiai, M. M., & Arif, M. (2024). Chat-GPT; validating Technology Acceptance Model (TAM) in education sector via ubiquitous learning mechanism. *Computers in Human Behavior, 154*, 108097. https://doi.org/10.1016/j.chb.2023.108097

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296.

Schiff, D. (2022). Education for AI, not AI for education: The role of education and ethics in national AI policy strategies. *International Journal of Artificial Intelligence in Education, 32*(3), 527–563. https://doi.org/10.1007/s40593-021-00270-2

Sethumadhavan, A. (2019). Trust in artificial intelligence. *Ergonomics in Design, 27*(2), 34–34. https://doi.org/10.1177/1064804618818599

Setzu, M., Guidotti, R., Monreale, A., Turini, F., Pedreschi, D., & Giannotti, F. (2021). Glocalx-from local to global explanations of black box ai models. *Artificial Intelligence, 294*, 103457. https://doi.org/10.1016/j.artint.2021.103457

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shkedi, A. (2004). Second-order theoretical analysis: A method for constructing theoretical explanation. *International Journal of Qualitative Studies in Education, 17*(5), 627–646. https://doi.org/10.1080/0951839042000253630

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist, 57*(10), 1380–1400. https://doi.org/10.1177/0002764213498851

Stackpole, B. (2019). AI ain't for everyone — who trusts bots, and why. M.I.T. Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/ai-aint-everyone-who-trusts-bots-and-why. Retrieved April 25th, 2025.

Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th International Learning Analytics and Knowledge Conference* (pp. 345–356). https://doi.org/10.1145/3576050.3576147

Torrecilla-Salinas, C. J., De Troyer, O., Escalona, M. J., & Mejías, M. (2019). A Delphi-based expert judgment method applied to the validation of a mature Agile framework for Web development projects. *Information Technology and Management, 20*, 9–40. https://doi.org/10.1007/s10799-018-0290-7

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–39. https://doi.org/10.1145/3476068

Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte., C, Spikol, D., Milrad, M; Cohelho, R., & Kizilcec, R. F. (2024). What Explains Teachers'

Trust in AI in Education Across Six Countries?. International Journal of Artificial Intelligence in Education, 1–29. https://doi.org/10.1007/s40593-024-00433-x

Villasenor Alva, J. A., & Estrada, E. G. (2009). A generalization of Shapiro–Wilk's test for multivariate normality. *Communications in Statistics—Theory and Methods, 38*(11), 1870–1883. https://doi.org/10.1080/03610920802474465

Wang, X., & Yin, M. (2022). Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems, 12*(4), 1–36. https://doi.org/10.1145/3519266

Wang, D., Bian, C., & Chen, G. (2024). Using explainable AI to unravel classroom dialogue analysis: Effects of explanations on teachers' trust, technology acceptance and cognitive load. *British Journal of Educational Technology, 55*(6), 2530–2556. https://doi.org/10.1111/bjet.13466

Williams, R., & Yampolskiy, R. (2021). Understanding and avoiding ai failures: A practical guide. *Philosophies, 6*(3), 53. https://doi.org/10.3390/philosophies6030053

Williamson, K., & Kizilcec, R. F. (2021). Effects of Algorithmic Transparency in Bayesian Knowledge Tracing on Trust and Perceived Accuracy. *In: Proceedings of The 14th International Conference on Educational Data Mining (EDM21). International Educational Data Mining Society*, pp 338–344.

Williamson, K., Kizilcec, R., Fath, S., & Heffernan, N. (2025). Algorithm Appreciation in Education: Educators Prefer Complex over Simple Algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 527–536). https://doi.org/10.1145/3706468.3706535

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2016). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience, 28*(14), 3866–3878. https://doi.org/10.1002/cpe.3745

Yudkowsky, E. (2023). Pausing AI developments isn't enough. *We need to shut it all down* (p 29). Time magazine. https://empoweruohio.nyc3.cdn.digitaloceanspaces.com/Pausing%20AI%20Development.pdf

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 1–27. https://doi.org/10.1186/s41239-019-0171-0

Authors and Affiliations

Yael Feldman-Maggor[1] · Mutlu Cukurova[2] · Carmel Kent[3] · Giora Alexandron[1]

✉ Giora Alexandron
giora.alexandron@weizmann.ac.il

Yael Feldman-Maggor
yael.maggor@gmail.com

Mutlu Cukurova
m.cukurova@ucl.ac.uk

Carmel Kent
kent.carmel@gmail.com

[1] Weizmann Institute of Science, Rehovot, Israel

[2] University College London, London, UK

[3] The Open University, Milton Keynes, UK