Received: 14 December 2021 | Accepted: 22 April 2022

DOI: 10.1111/bjet.13232

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

Teachers' trust in AI-powered educational technology and a professional development program to improve it

Tanya Nazaretsky[1] | **Moriah Ariely**[1] | **Mutlu Cukurova**[2] | **Giora Alexandron**[1]

[1]Department of Science Teaching, Weizmann Institute of Science, Rehovot, Israel

[2]UCL Institute of Education, University College London, London, UK

Correspondence
Tanya Nazaretsky, Department of Science Teaching, Weizmann Institute of Science, Herzl St 234, Rehovot, Israel.
Email: tanya.nazaretsky@weizmann.ac.il

Funding information
Azrieli Foundation; The Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center

Abstract

Evidence from various domains underlines the critical role that human factors, and especially trust, play in adopting technology by practitioners. In the case of Artificial Intelligence (AI) powered tools, the issue is even more complex due to practitioners' AI-specific misconceptions, myths and fears (e.g., mass unemployment and privacy violations). In recent years, AI has been incorporated increasingly into K-12 education. However, little research has been conducted on the trust and attitudes of K-12 teachers towards the use and adoption of AI-powered Educational Technology (AI-EdTech). This paper sheds light on teachers' trust in AI-EdTech and presents effective professional development strategies to increase teachers' trust and willingness to apply AI-EdTech in their classrooms. Our experiments with K-12 science teachers were conducted around their interactions with a specific AI-powered assessment tool (termed AI-Grader) using both synthetic and real data. The results indicate that presenting teachers with some explanations of (i) how AI makes decisions, particularly compared to the human experts, and (ii) how AI can

Tanya Nazaretsky and Moriah Ariely contributed equally to this study.

complement and give additional strengths to teachers, rather than replacing them, can reduce teachers' concerns and improve their trust in AI-EdTech. The contribution of this research is threefold. First, it emphasizes the importance of increasing teachers' theoretical and practical knowledge about AI in educational settings to gain their trust in AI-EdTech in K-12 education. Second, it presents a teacher professional development program (PDP), as well as the discourse analysis of teachers who completed it. Third, based on the results observed, it presents clear suggestions for future PDPs aiming to improve teachers' trust in AI-EdTech.

Practitioner notes

What is already known about this topic

- Human factors, and especially trust, play a critical role in practitioners' adoption of technology.
- In recent years, AI has been incorporated increasingly into K-12 education.
- Little research has been conducted on the trust and attitudes of K-12 teachers towards the use and adoption of AI-powered Educational Technology.

What this paper adds

- This research emphasizes the importance of increasing teachers' theoretical and practical knowledge about AI in educational settings to gain their trust in AI-EdTech in K-12 education.
- It presents a teacher professional development program (PDP) to increase teachers' trust in AI-EdTech, as well as the discourse analysis of teachers who completed it.
- It presents clear suggestions for future PDPs aiming at improving teachers' trust in AI-EdTech.

Implications for practice and/or policy

- Pre- and in-service teacher education programs that aim to increase teachers' trust in AI-EdTech should include a section providing teachers with a basic understanding of AI.
- PDPs aimed to increase teachers' trust in AI-EdTech should focus on *concrete* pedagogical tasks and specific AI-powered tools that are considered by teachers as helpful and worth the effort to learn.
- AI-EdTech should not restrict teachers to follow specific pedagogical scenarios, but rather provide teachers with the freedom to design and implement various types of pedagogies that meet their preferences, students' needs, and classroom reality.
- Teacher agency is key to gaining their trust. AI-EdTech should allow teachers to review, modify, and if necessary, override AI-based recommendations before they are sent to students.

INTRODUCTION AND LITERATURE REVIEW

There are few studies undertaken on whether and how K-12 teachers' data literacy in general and AI knowledge in particular influence their trust in AI-EdTech and potential resistance to integrating AI-EdTech in their practice. Well-established theories such as the theory of academic resistance (Rienties, 2014), the resistance to organizational change theory (Piderit, 2000), and the technology acceptance model (TAM) (Davis, 1989) treat the issue of practitioners' perceptions and attitudes towards using technology and resistance to a change related to technology adoption in general. However, none of the above theories focuses on the specific features of AI-based tools. In contrast to EdTech in general, AI-powered EdTech focuses on computer systems capable of accomplishing tasks and activities that have historically relied on human cognition (Brown et al., 2020), such as natural language processing and reasoning. Previous research indicates that the aforementioned "AI-nature" of technology plays a significant role in forming practitioners' perceptions and attitudes towards adopting AI-based EdTech.

For instance, in a large sample study, previous research has presented some misconceptions and negative perceptions related to the AI-nature of AI-EdTech (Cukurova et al., 2020). Moreover, research in other domains confirms that human forecasters quickly lose trust in automated recommendation systems after seeing that they make a mistake, while they are more tolerant of the same mistake made by a human (Dietvorst et al., 2015). This phenomenon is called *Algorithm Aversion* (Dietvorst et al., 2015) and manifests itself also in educational settings. More specific to teachers, it was shown that they may expect automated recommendations to be fully compliant with their own opinion and perceive the recommendation worthless in the case of any disagreement. Such possible differences *between AI-EdTech suggestions and teachers' opinions* are likely to reduce their trust in AI-EdTech (Nazaretsky et al., 2022). These phenomena might be attributed to the manifestation of fundamental human bias and heuristics of judgement under uncertainty (Tversky & Kahneman, 1974), such as *Confirmation Bias* and *Belief Perseverance* (Nickerson, 1998), and they can further lead teachers' resistance to the adoption of AI-EdTech recommendations (Nazaretsky et al., 2021).

In the relevant field of learning analytics, in higher education settings, Tsai and Gasevic presented six challenges related to the adoption based on a comprehensive review of 23 empirical studies of learning analytics. One has to do with the end-users lack of ability to understand learning analytics and interpret the presented data (Tsai & Gasevic, 2017). Research confirmed that to use learning analytics effectively (i.e., critically interpret the results rooted in big data analysis and make data-informed decisions), end-users need to have basic data literacy (Wolff et al., 2016). In addition, knowing how the data is collected, stored and processed in AI-powered systems is critical in developing students' and instructors' trust in AI-EdTech (Pardo & Siemens, 2014). On par with these findings from higher education settings, our previous research from K-12 settings indicates that although teachers consider AI-EdTech as a potentially powerful means for improving their instruction, they might avoid incorporating it into their day-to-day teaching routines. Among the reasons for this resistance is teachers' *lack of knowledge about AI* (e.g., how AI-EdTech learns and makes decisions and what kind of information it can provide) (Nazaretsky et al., 2021, 2022).

Nevertheless, emerging evidence suggests that training programs and interventions can help mitigate AI-related biases and improve decision-making processes (Morewedge et al., 2015; Sellier et al., 2019) by increasing practitioners AI knowledge. For example, there are calls to "reboot" medical education to adjust it to the age of AI in the medical domain by teaching medical students the internals of machine learning and AI (Wartman & Combs, 2018). In higher education settings, research shows that relevant training programs aimed to raise end-users' understanding of AI and learning analytics are crucial in

adoption success (Goldstein & Katz, 2005). Moreover, it has also been shown that scaffolding faculty by personal consultants while working with data provided by AI-EdTech improves their ability to interpret the presented insights and confidence in building follow-up data-driven pedagogical decisions (Arnold et al., 2014). Based on these considerations and evidence, we hypothesize that training practitioners can positively affect their attitudes towards adopting AI-powered technology. However, *what exactly practitioners should know* about AI to understand the nature of AI, to be able to explain and justify the origin of AI decisions, and to use AI's results in their practice effectively is still an active research area (Grunhut et al., 2021; McCoy et al., 2020). We aim to contribute to this literature with a professional development program (PDP) designed to improve K-12 teachers' knowledge of AI and trust in AI-EdTech.

Professional development program and research rationale

The PDP developed in this study focused on a specific application of Natural Language Processing (NLP) and AI for automated grading of constructed responses in biology (Ariely et al., 2022), and had two main goals: (i) to promote teachers' understanding of causal explanations in biology and how to teach and assess written explanations, using analytic grading rubrics; and (ii) to promote teachers' understanding of AI-EdTech, including affordances and limitations of automated assessment in comparison to human assessment. While this study centres on the second goal, these two goals are well-connected. The assessment design based on the analytics rubric prepared the ground for discussing how AI can be used to automate the scoring process. We hypothesized that increasing teachers' knowledge about AI and how it makes decisions, as well as providing them opportunities to experience it by working alongside AI-EdTech in their authentic environment, would lead to a more nuanced understanding of the capabilities and limitations of Human Agency and AI-EdTech, improve their attitudes towards using AI-EdTech, and in turn would increase their trust in AI-EdTech.

Trust is rarely defined in studies of AI-assisted decision-making (Vereschak et al., 2021), and educational contexts are no exception. In this study, we adopted the definition proposed by Lee and See that “trust 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, p. 54). Aligned with the elements of trust identified from a recent literature review of the field (Vereschak et al., 2021), our PDP links trust to a situation of *vulnerability, positive expectations*, and considers it as an *attitude*. More specifically, we developed a PDP that created an appropriate experimental environment suitable for studying trust (Subsection “*Experimental protocol for studying trust*” explains the uncertainty of the outcomes of the decisions taken), focused on influencing negative factors decreasing teachers' trust (Subsection “*Factors affecting trust*”), and provided us with an opportunity to monitor teachers' *perceptions* through teacher group discussion and study *attitudes* towards using AI-EdTech in their pedagogical practice.

Research goals and questions

Thus, the goal of the present study was to learn about teachers' perceptions and attitudes towards using AI-EdTech, whether they change during the PDP, or not, and how such potential changes manifest themselves in teachers' trust. This was formulated through the following research questions:

RQ1. To what extent does teachers' knowledge about AI-powered assessment change throughout the PDP?

RQ2. To what extent do teachers' perceptions and attitudes towards human and AI-powered assessment change throughout the PDP?

RQ3. Whether the above potential changes reflect in teachers' trust and willingness to adopt AI-EdTech?

METHODOLOGY

Research context and population

The context of this study was an eight-week, sixteen-hour unit that was part of a thirty-hour PDP for in-service high-school biology teachers, held at the Weizmann Institute of Science, Israel.

Six in-service high school biology teachers participated in the PDP. The teachers taught in public schools from various locations in Israel. Their teaching experience ranged from one year to more than ten years (see Table 1).

AI-Grader for automated formative assessment

During the PDP, the teachers were presented with AI-Grader—an NLP and AI-powered assessment technology that analyzes constructed responses (CR) to open-ended questions in Biology and generates an automated score based on an analytic grading rubric (Figure 1). The analytic grading rubric represents the required chain of reasoning, including all the main content and causal relations that should appear in a correct response (for a full description of the items, grading rubrics and the NLP-based algorithms, see Ariely et al. (2022).

Intervention design overview

To create the proper experimental environment for studying trust, we put our participants in *a vulnerable* position (as without vulnerability there is no need for trust to emerge [Castelfranchi & Falcone, 2010]) and established initial *positive expectations* regarding AI-Grader (as trust can only be formed if the negative outcomes related to trust is considered very unlikely [Lewis & Weigert, 1985]). In the next sections, we describe what

TABLE 1 The teachers' teaching experience

Teacher (pseudonyms)	Teaching experience (range in years)
Betty	<5
Usher	<5
Nicole	6–10
Jenny	>10
Danna	>10
Miriam	>10

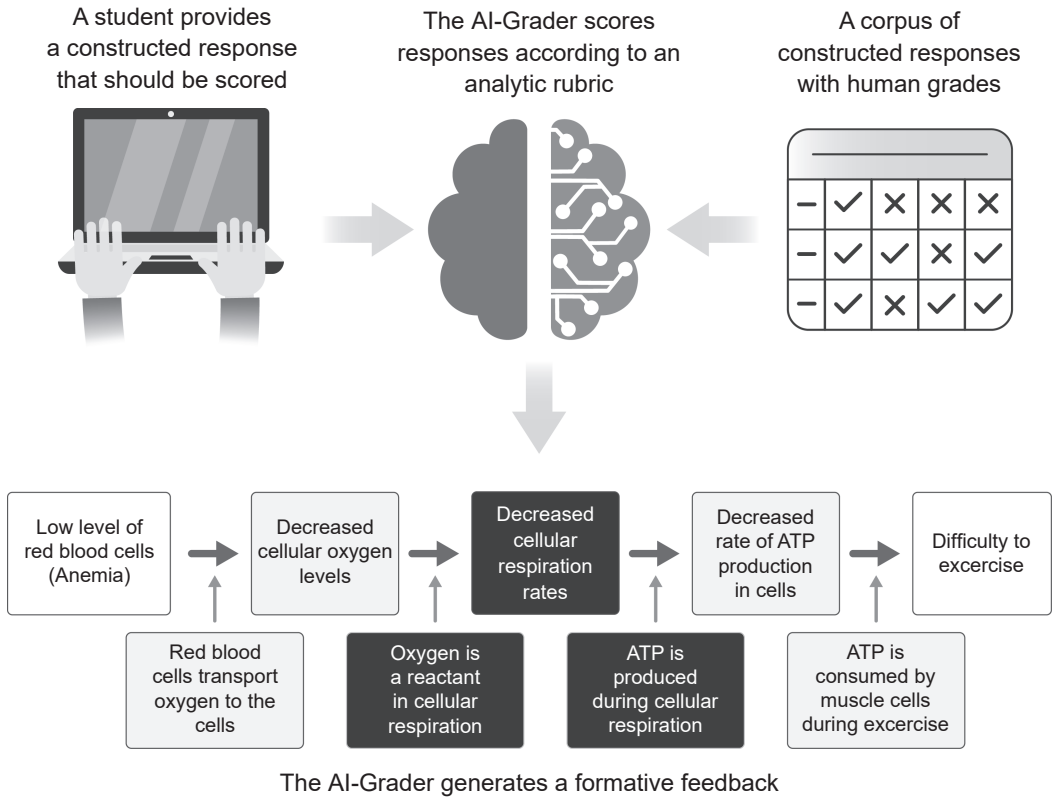

FIGURE 1 The workflow of the tool for automated grading of constructed responses. The key components of the full chain of reasoning that should be present in a correct answer are presented in the flow diagram (bottom of the figure). The specific example presents the feedback on a certain type of conceptual error that was provided to the students as infographics. White squares represent the information given in the question. The light grey squares indicate that the student correctly mentioned the corresponding property, while the black squares indicate that the student missed the corresponding part of the written explanation

factors that influence trust we chose to monitor, and how the PDP was explicitly designed to impact these factors.

Experimental protocol for studying trust

Vulnerability occurs when there is uncertainty in the outcomes of a decision that genuinely matters to the teachers. This uncertainty might be due to the unpredictable nature of the context or the lack of knowledge/skills of the decision-maker. In this study, we presented the teachers with automated reports based on *their own students' constructed responses* and asked them to use the reports in their classrooms as part of their usual pedagogical routine. This context created a situation of vulnerability as teachers had serious incentives and their real students at stake.

Positive expectations are another key element of trust. To establish *initial* positive expectations, and in line with (Yin et al., 2019), we presented the teachers with the accuracy of an ML automated grading system that was trained and tested on a large corpus (two open-ended questions, ~1300 constructed responses from ~650 students, ~85% accuracy). Such level of accuracy is reasonable in the context of the CR assessment task, so communicating it formed a fair level of positive expectations before teachers experienced AI-Grader's results.

Factors affecting trust

During the PDP, our goals were to influence: (i) teachers' lack of AI knowledge; (ii) misconceptions related to human "super-powers" and to expectations that AI-Grader's recommendations should always be perfect (which the teachers tended to interrpet as "fully compliant with my opinion" [Nazaretsky et al., 2021]).

1. *Lack of AI knowledge*: Our goal was to improve the participating teachers' ML and AI literacy, and not to turn them into AI experts. Thus, we did not overload the teachers with technical details and jargon. Instead, we exposed them to some of the concepts, standard procedures and common practices of ML (i.e., procedural knowledge; OECD, 2019). We taught only procedures that the teachers could fully understand based on their prior knowledge, such as rubric development, expert tagging, accuracy, etc. We hypothesized that this knowledge would allow the teachers to understand and discuss the rationale for common ML practices. Specifically, we wanted to ensure that the teachers would be able to interpret and justify automated results and communicate them to the students by the end of the PDP.
2. *Misconceptions*: Human misconceptions are hard to tackle, so we developed a creative idea inspired by "The Masked Singer" television show (the show was broadcasted on a leading TV channel during the PDP and was quite popular). We called it "The Masked Rater" activity. First, we asked the teachers to grade the same 30 students' CR. Based on the results, we calculated the agreement rates between the teachers themselves, the teachers and the expert, and the teachers and AI-Grader. Then, we gave each teacher an individual report containing her/his agreement rates with four "masked colleagues (these were two other participating teachers, the expert, and AI-Grader), however, we did not tell them who those colleagues are. All the agreement rates ranged between 79% and 91%. At this point, the teachers came to realize that their average agreement *with all raters* was around 85%. Then, the teachers discussed why the grades might vary and thought about potential reasons for giving different grades to the same constructed response. Finally, we revealed to the teachers that one of the "masked colleagues" was AI-Grader.

This activity, we hypothesized, would give teachers a perspective of the expected rate of disagreement between peers on the task of grading CRs. It emphasized that a certain degree of disagreement is normal even among human graders and that it is unrealistic for one to expect to always have a full agreement with AI-Grader's results due to the fact that grading open-ended questions involves some level of subjective interpretation.

The professional development program structure

The PDP included three group sessions (~9 hours), and teachers' independent, hands-on assignments (~7 hours), as detailed in Figure 2. In addition, we conducted one-on-one interviews (~2 hours with each participant separately). The AI-Grader's results for the interviewee teacher's own class were presented, explained, and discussed during each interview. Teachers proposed related follow-up activities that they later conducted with their classes. However, the interviews mainly addressed the first goal of the PDP (namely, how to teach and assess written explanations, using analytic grading rubrics).

Since the teachers marked the "The Masked Rater" activity as an important triggering event that played a key role in changing their perception of interpreting AI-Grader's recommendations, we refer to it as the cutting point between parts A and B of the course (see Figure 2). During the group discussions, we asked the teachers to argue the pros and

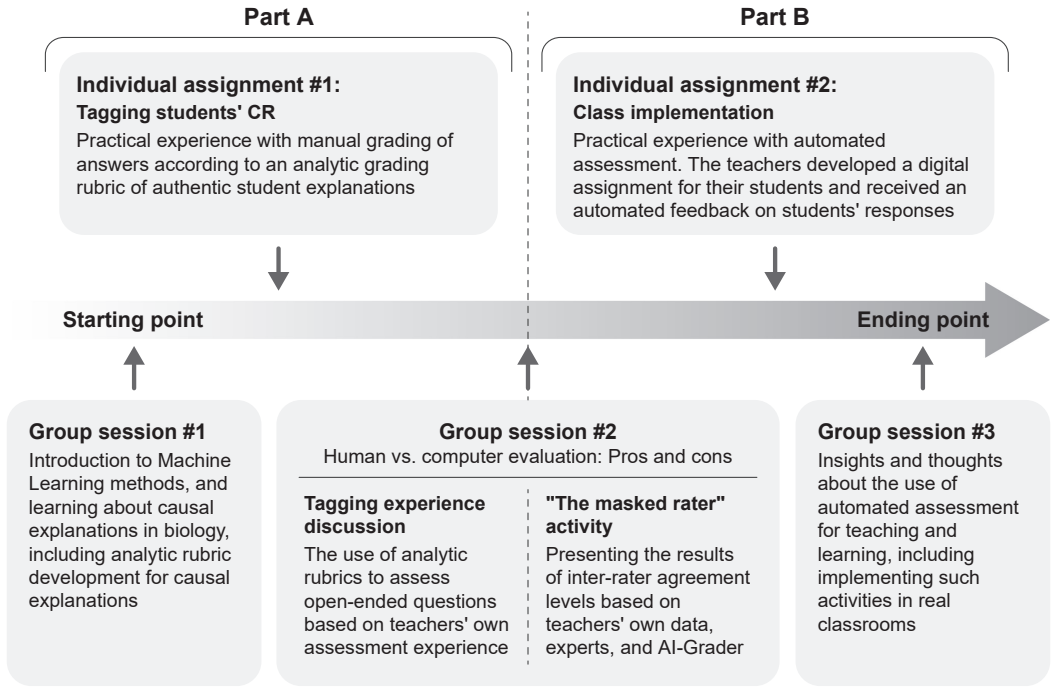

FIGURE 2 Timeline of the group sessions and assignments throughout the PDP. The first part of the PDP (part a) included group session #1, the tagging assignment and half of group session #2. The second part of the PDP (part b) included half of group session #2, the implementation assignment and group session #3

cons of automated assessment of causal explanations based on analytic rubrics vs. human teacher holistic assessment: (i) first time (during Group session #1) based on a theoretical presentation of the assessment method based on the analytic rubric and how AI-Grader implements it; (ii) second and third time (during Group session #2) the teachers reflected on their experience to manually apply the method with a synthetic class before and after "The Masked Rater" activity; finally, (iii) fourth time (during Group session #3) after they experienced the method in their real classes.

Data collection and discourse analysis

The PDP was conducted during the 2020–21 academic year (online via Zoom due to the COVID-19 pandemic). All sessions were recorded and transcribed, and the transcripts were analysed in an iterative process.

We used a bottom-up approach to analyse all the relevant group discussions to identify the categories and subcategories that emerged from the teachers' discussions (see Table 2). The unit of analysis was any segment of the discourse where teachers elicited their knowledge, perceptions or attitudes towards AI-EdTech or human assessment. The data analysis was conducted by the first two authors, and the utterances were classified into categories as followed. First, the authors worked together to classify ~10% of the data. This stage allowed us to clarify and refine the classification of data into categories. Next, the authors analysed 40% of the data independently, and the inter-rater agreement was computed by calculating the accuracy rate. The agreement between the raters on the mapping of sentences into categories at this stage was ~82%, and the majority of disagreements were related to compound teacher sentences that addressed two different

TABLE 2 Emerged categories from the discourse analysis ($N = 125$)

Categories (utterances)	Sub-categories	Categories descriptions
Knowledge about AI ($N = 36$)	(i) Procedural knowledge ($N = 19$) (ii) Epistemological knowledge ($N = 17$)	Utterances in which the teachers either elicited their prior knowledge about AI and ML or asked for this knowledge during the PDP This included (i) procedural ML knowledge, such as rubric development, the manual tagging process; and (ii) epistemological knowledge about the nature and limitations of automated assessment, such as the significance of creating high-quality training data, why the results are highly context-specific and when they can or cannot be extended to other contexts
Perceptions and attitudes towards the human grading and automated process ($N = 65$)	(i) Accuracy & reliability ($N = 39$) (ii) Technical issues ($N = 22$) (iii) Affective ($N = 4$)	Utterances in which teachers talked about or revealed their attitudes towards (i) accuracy and reliability of human and automated assessment; (ii) technical issues related to human and automated assessment, such as the time needed for assessment, tiredness, etc.; and (iii) their feelings and emotions towards AI tools
Blending AI recommendations into pedagogy ($N = 24$)	–	Utterances in which the teachers talked about their thoughts about incorporating AI-Grader in their pedagogy

categories. After discussions, the unit of analysis and classifications were refined again (e.g., compound sentences were divided into two or more units of analysis), resulting in an increase in the inter-rater agreement (~95% accuracy rate). Finally, the remaining data were analysed by the first author. In total, we collected and mapped 134 teacher utterances from the three group session discussions. In the results section below, we elaborate on the emerged categories and subcategories and provide examples of the utterances and their mapping.

RESULTS

The bottom-up analysis revealed three major categories in teachers' discussions about AI in general and AI-Grader specifically: (i) knowledge about AI and AI-Grader; (ii) perceptions and attitudes towards the human grading and automated grading process; and (iii) perceptions and attitudes towards blending AI-Grader recommendations into pedagogy. Irrelevant utterances that did not fit any of the above categories were categorized as 'Other' ($N = 9$) and were omitted from the total amount of utterances. Thus, we analysed a total of 125 utterances. Elaboration on each category, and its sub-categories, is provided in Table 2.

RQ1: Acquisition of AI-related knowledge

During the course, the teachers asked many questions to clarify their understanding of AI-Grader, and related procedures. In the first part of the course (part A), we mostly provided the teachers with procedural knowledge about the AI-powered assessment process. As expected, most teachers' utterances in the first part of the PDP were statements and questions about ML procedures (i.e., procedural knowledge, Figure 3). However, at the very beginning, the teachers' elicited knowledge and questions were general, superficial and naïve:

> Danna: "This machine, this software, is it something that learns?"
>
> Nicole: "What is the output that I can get [from the system] as a teacher?"

As the teachers acquired more knowledge about and experience with AI-powered assessment, they started to ask more concrete, accurate and advanced questions about ML procedures:

> Miriam: "How many graded responses are needed to achieve such [high] level of agreement between the computer and the human graders?"
>
> Nicole: "So, the system learns from experts, it knows only what it was trained for. However, can it learn additional stuff if it is provided with properly graded examples?"
>
> Danna: I thought about: the more answers and more such […] examples, then the more it [AI-Grader] will know?"

In the second part of the course (part B), the majority of utterances included statements and questions about the nature of AI-powered assessment and AI in general, and about their affordances and limitations (i.e., epistemological knowledge, Figure 3):

> Danna: "If the AI works according to some kind of rubric, and I know that someone developed it [the rubric], it is something that experts developed. I have no reason not to agree with it …"
>
> Usher: "We need to learn how to let go. It is OK if [AI-Grader] makes errors every once in a while, everyone makes mistakes. Even the computer is human because it is based on people [training it] …"
>
> Jenny: "It's like the student wrote down all the "right" words, but I can see that he doesn't understand anything … and I thought to myself that this is so annoying!

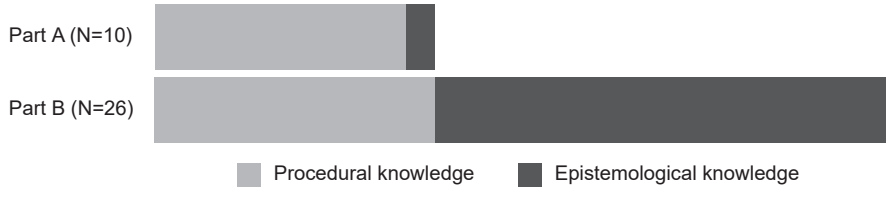

FIGURE 3 Teachers elicited knowledge and questions on AI-grader and AI in general, in the first and second parts of the professional development program ($N = 36$)

> If AI-Grader were to grade this answer, it would give a really high score, because seemingly all the components of the answer are there …"

As presented in the examples above, we found a shift in teachers' discourse regarding their knowledge about AI-Grader, as revealed from their statements and questions throughout the course (Figure 3).

As presented in Figure 3, the teachers asked more questions and provided more statements in the second part compared to the first part of the course. However, more importantly, the discourse analysis showed that the teachers' knowledge evolved from a superficial, naïve understanding of AI-EdTech, to a deeper, more sophisticated, and advanced understanding. This understanding may have allowed them to think more critically and practically on AI-Grader and argue about its possible application in their pedagogy as we elaborate on in the following sections.

RQ2: Teachers' perceptions and attitudes towards human and automated grading processes

Regarding the teachers' perceptions and attitudes towards human and automated grading, we found that teachers mostly talked about issues of accuracy and reliability and also about technical and affective issues regarding the grading process.

When we asked the teachers how the grading experience was, evidence from the discourse analysis confirmed that human assessment of constructed response items is challenging and even overwhelming to some teachers, as well as extremely time-consuming:

> Danna: "At first it was really difficult even to start [grading]. I saw 30 answers, and I said to myself- this reminds me of an exam. I am putting this away. And then, for each answer I went back to the rubric again and again. Then I took two computers, then I tried another strategy because it just wasn't ending …"

> Usher: "I must admit, that for me it was quite a nightmare. I'm sorry. At some point, I was on the verge of despair, and I started saying to myself OK, ten more … you can do it, just nine more, you can do it. And writing the feedback … I admit that at some point there were some students I skipped."

> Miriam: "It took me a lot of time … it took me hours."

Some of the teachers also perceived human grading to be non-100% accurate and less reliable in many cases:

> Jenny: "I had a hard time scoring some of the responses. Is it [the category] there? Is it not there?"

> Usher: "I expect to see [in students' responses] some kind of thinking that is something close to what I think … now if they wrote it a little differently, then I, as their teacher, can miss it …"

> Nicole: "… many times you read an answer that you know is not accurate, and it is very hard to decide how many points to take off. It is all mixed up, with no criteria."

Despite the difficulties, the teachers perceived their ability to grade students' CRs as superior to the computer's ability to grade the same CRs. However, evidence from the discourse analysis revealed two interesting phenomena in this respect: *confirmation bias* and *algorithm aversion*.

The discourse about the grading process provided strong evidence that the teachers have a strong confirmation bias that humans are superior to AI in this grading task. All the teachers argued against the automated assessment because it was considered as incapable of seeing "the big picture" or "reading between the lines" as humans can. Namely, the teachers said that they "know what a student knows" even if he or she did not write it explicitly in their response:

> Miriam: "Sometimes things are implied from the student's response, from his response it implies that he understands [the content]. The algorithm can't do this, either there is [the category in the response], or there isn't. But I can understand that he understands. So, I do give him some points, not all, but … I believe that the algorithm can't do that …"

One of the implications of this confirmation bias is that the teachers are influenced by their previous knowledge of the students when analysing their responses, leading to subjectivity. This was also manifested in how teachers treated differently responses of students that were not theirs (thus, they did not have prior knowledge of them):

> Danna: "At first I assessed as a teacher, but then I thought- these are not my students, I can assess like a machine: [the category is] present or not present [in the response]."

> Jenny: "At some point I said to myself, wait, there is no student behind the scenes. I will be cold and correct."

Regarding the *algorithm aversion*, some teachers perceived the possible machine errors as not acceptable ("perfect or worthless" perception) or less acceptable than human errors:

> Nicole: "I have a dilemma because I feel I can't trust it 100%."

> Miriam: "In this case it is better [to grade CRs] on our own."

> Usher: "Still, with a fellow teacher - I feel more comfortable."

After the teachers were confronted with the fact that human-to-human inter-rater disagreements can actually be quite significant ("The Masked Rater" activity), and the fact that they may have a higher agreement with the algorithm than with their peers, we observed a change in the teachers' discourse. There was a noticeable reduction in the frequency of teachers' utterances that revealed confirmation bias and algorithm aversion (Figure 4).

At the same time, new considerations emerged including "not perfect is not worthless,"

> Usher: "… I feel that it requires us teachers to do something that is always difficult for us to do - to let go. And it's like saying okay, it is perfect, no it is not perfect, … just trust and send."

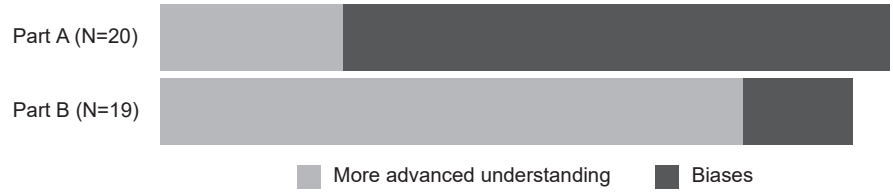

FIGURE 4 Teachers' perceptions and attitudes towards the human and AI grading process (accuracy and reliability), in the first and second parts of the professional development course ($N = 39$). Biases (dark grey) included expression of biases related to human superpowers and algorithm aversion, while more advanced understanding (light grey) includes the perception that AI-grader can be more reliable and accurate than humans and its possible errors are OK

> Betty: "I can say that the table you showed us with the agreement between us, and with the expert, forget about the algorithm … even between us, I mean … the bottom line is that if it [the agreement] is between teachers, or between teachers and an expert, or teachers with the algorithm, it is [the agreement] around 85%. For me it says a lot."

and "computers can be more reliable and accurate than humans":

> Nicole: "I was excited about it [automated grading], about being able to evaluate [the responses] objectively. For me personally, I think, it is hard [to do] … I think it [automated grading] can point- better than the teacher, which might be influenced from her personal acquaintance with the students, to the students' weaknesses that should be strengthened. We don't always succeed in pointing at them."

> Betty: if you have 120 exams to grade, you have a greater chance to miss something, than with the help of the computer."

> Usher: "As much as I try to be objective, I speak for myself, there are certain names. I look at who wrote, even the handwriting influences my grading. Although this shouldn't happen, I'm human."

RQ3: Teachers' trust and willingness to use AI-EdTech in pedagogy

During the PDP, the teachers discussed many times their ability and willingness to blend AI-Grader into their pedagogy. In the beginning, the teachers talked less about their pedagogy. This is not surprising since they had little knowledge or experience with AI-EdTech. However, during the course, due to the newly acquired knowledge and experience with AI-Grader, the teachers were able to think about if and how to incorporate such tools into their pedagogy. In parallel to the above-mentioned shift in the teachers' discourse about accuracy and reliability, in the second part of the PDP, the teachers also expressed positive attitudes regarding AI-Grader and were able to think about how and when it can help them in their pedagogy:

> Usher: "… I'm sure that if I will look at the exams that I grade when I have a stack of exams to grade, I have a fatigue curve …"

> Nicole: "I had a full picture of my class, where do they stand, what are their problems, as a class and also on each student. On the other hand, it didn't save me time because I wrote each student a personalized feedback."
>
> Betty: "I think it [seeing the full class picture] will happen faster. For us, it takes a lot of time, especially with a new class, it takes time until you get to know them … this [automated grading] is fast and reliable …"

In the last session, we asked the teachers if and how they see AI-Grader integrated into their pedagogy. At this point, the teachers were able to think of various ways of using the automated grading, which was based on the more advanced knowledge and perceptions, as previously described. Below, we provide a few examples of teachers' suggestions on how and when to blend AI-Grader's recommendation into their pedagogy:

> Danna: "I will check the answers because there are responses that the computer classifies in some way, but maybe I didn't teach [this content] or didn't emphasize it, so I can consider that [when looking at the automated grading]."
>
> Betty: "while I was doing this [tagging] I thought all sorts of stuff. One was that it is very important that all these components (of the grading rubric] will be considered. So, it is good that a computer will do this. That way I can read what the computer did and analyzed, and I can complete it, like completing the feedback."
>
> Danna: "I can choose the questions that most of the students didn't get right in many of the rubric components. These are the questions that I will use for getting my conclusions [about the class]. And when I look at it for per student, I can teach in small groups about these specific problems …"
>
> Nicole: "I think it is easy for us when it comes to the extremely good or the extremely struggling students. Those [students] in the middle, are in the middle because they didn't comprehend A, or maybe they didn't comprehend B. It is very difficult to make these groups [of students with a specific similar error], I think this tool can help."
>
> Usher: "I feel that this [automated scoring] allows me to create a language with my students. To show them what they did [in their response]. For me to do it, it will be complicated, and I am willing to pay some price of inaccuracy of a few points for that. I don't mind …"

DISCUSSION AND CONCLUSIONS

Research in various domains (e.g., healthcare, transportation, etc.) highlighted the critical role that human factors play in the adoption of technology. Especially, users' perspectives and attitudes shape their trust, which is vital to their willingness to follow and accept automated recommendations (Buckingham Shum et al., 2019). In the case of AI-EdTech, the issue is even more complex due to practitioners' AI-specific misconceptions and lack of a general understanding of how AI-EdTech operates (Brynjolfsson & McAfee, 2014; Cukurova et al., 2020). The picture is not different in the K-12 educational context, and the adoption of AI-EdTech can be significantly slowed down by teachers' lack of trust and negative attitudes. This study is a part of our continuous effort to shed light on K-12 teachers' perceptions and

attitudes towards adopting AI-EdTech and investigate ways to improve teachers' trust. In this study, we focused on designing an effective PDP to increase teachers' trust and their willingness to adopt AI-EdTech.

The design of the PDP meets the recommendations of Vereschak et al. (2021) to create an experimental setting that is suitable for measuring *trust* as an attitude that expresses itself in a situation of vulnerability and in the presence of initial positive expectations. In efforts to support and study teachers' trust in AI-EdTech, it is paramount that the PDPs create conditions that are conducive to generating trust. Working in such an environment, we chose to focus on several aspects that are known as negative factors influencing trust, such as teachers' lack of knowledge about AI-EdTech and misconceptions related to AI-EdTech. Our results confirmed that during the PDP, teachers gained important knowledge about AI-powered assessment (exemplified by AI-Grader). As mentioned above, AI-Grader is a tool for automated formative assessment of constructed responses. As such assessment task requires critical thinking, logical reasoning, and understanding of texts written in a free language, it is usually associated with human-only capabilities and serves our purpose of exemplifying the “AI nature” of AI-EdTech in general. The gained AI knowledge helped teachers to mitigate some of their misconceptions and biases related to the AI-nature of AI-EdTech. Besides, the extended knowledge about AI allowed the teachers to correctly interpret AI-Grader results and to justify its potential mistakes or inconsistencies with their own opinion. Thus, by the end of the PDP, the teachers expressed their willingness to use AI-Grader in their classrooms. Moreover, they were able to propose innovative ways to integrate AI-Grader into their pedagogy and use it for making data-driven decisions, reinforcing the authenticity of their positive attitude.

We attribute this potential attitude change to the increase in their level of trust in AI-Grader in specific, and in AI-EdTech in general.

Recommendations for PDP designers and AI-EdTech developers

Based on the evidence presented in the paper, we suggest PDP creators include the following key elements in their PDPs.

First, we propose to focus on a *concrete* pedagogical task and AI-powered tool that is considered by teachers as helpful and worth an effort to learn. Moreover, the usefulness of the tool is better to be presented in contrast to humans' ability to perform a similar task and/or the effort required to complete the task. For example, in our case, teachers can manually grade constructed responses; however, it is a highly time-consuming and exhausting task, so the value proposition of AI-Grader is clear. Moreover, we suggest (in cases where there are no privacy and ethical concerns involved) using real participant data in experimental settings. This allows giving teachers an opportunity to evaluate the presented AI-powered tool's usefulness and easiness to use, which are known as the important factors influencing teachers' willingness to adopt technology in general (Davis, 1989). In addition, the use of data from real students creates a situation of vulnerability as teachers had a genuine stake in AI-EdTech decisions.

Second, the proposed PDP should include a section providing teachers with a basic understanding of AI-common procedures and practices. Although we cannot make all teachers AI experts and the question of how much technical knowledge is enough for trust is still open (Kizilcec, 2016), the PDP should allow teachers to effectively judge the validity of the automated results and give them a perspective of what kind of errors the system can make. For example, in our case, we emphasized that the implication of learning from graded examples is replicating possible biases in human gradings, and the system's knowledge is restricted to what phenomena exist in an input corpus. This recommendation is consistent

with procedural justice theory (Lind & Tyler, 1988), which states that providing procedural transparency (in our case, the knowledge of what rubrics were used for grading, and how ML systems learn based on graded examples) makes people more tolerant to possible errors if they consider a procedure as just and fair.

Finally, it is very important to introduce teachers to the expected accuracy in the context of the presented task (e.g., by presenting the levels of human disagreement on the corresponding task, which is usually considered as a benchmark for ML accuracy). This knowledge can promote developing a more profound understanding that sometimes there is no single truth. Meaning, similarly to how teachers may differ in their interpretation of complex pedagogical scenarios, without this being interpreted as one of them has to be incorrect, disagreements can occur between human and machine gradings. More specifically, we designed activities pretending AI-EdTech to be one of the colleagues (e.g., "The Masked Rater" activity) as potential triggers to make teachers consider a recommendation originated from AI-EdTech as equivalent to advice from their peer. Thus, potentially reducing the negative effects of Confirmation Bias and Algorithm Aversion.

On the AI-EdTech design side, to reduce teachers' anxiety about mistakes that can occur in automated recommendations, our study emphasizes the importance of giving teachers the control to review and overwrite the recommendations before sending them to their students. Clear explanations of such opportunities that highlight human agency in AI-EdTech's use can significantly improve teachers' trust in their adoption. Therefore, should be considered as part of PDPs. In addition, in our study, teachers proposed various ways to use the automated results in their practice, which is on par with previous research considering blended pedagogy as a continuum rather than one specific scenario of using online technology in bricks-and-mortar settings (Powell et al., 2015). This pointed out the importance of designing AI-EdTech systems in which UI/UX does not restrict teachers to a specific pedagogical scenario, but rather provides freedom of choice based on their pedagogical preferences.

LIMITATIONS

Although our results are based on a self-chosen population that most likely has initial positive expectations towards using AI-EdTech, we observed in the participants the manifestation of several negative phenomena on par with the results from previous research based on a much larger and more diverse teachers' population (Nazaretsky et al., 2022). So, although our results and conclusions are based on the in-depth discourse analysis of a limited number of teachers, and on teachers working with a particular example of AI-EdTech that focuses on assessment (AI-Grader), they have the potential to be applied to a more diverse and general teachers' population and are worth to be explored in future studies.

ACKNOWLEDGEMENTS

We wish to thank the biology teachers who participated in the PDP. We thank Ziv Ariely for graphic design. TN is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. The work of MA was partially supported by the Israeli Council for Higher Education (CHE) via the Weizmann Data Science Research Center.

CONFLICTS OF INTEREST

There are no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data from this study are not available online due to privacy issues.

ETHICS STATEMENT

Approval for conducting this research was received from the Weizmann Institutional Review Board (IRB) for Education.

ORCID

Tanya Nazaretsky https://orcid.org/0000-0003-1343-0627
Moriah Ariely https://orcid.org/0000-0001-6539-3544
Mutlu Cukurova https://orcid.org/0000-0001-5843-4854
Giora Alexandron https://orcid.org/0000-0003-2676-6912

REFERENCES

Ariely, M., Nazaretsky, T., & Alexandron, G. (2022). Machine learning and Hebrew NLP for automated assessment of open-ended questions in biology. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00283-x

Arnold, K. E., Lynch, G., Huston, D., Wong, L., Jorn, L., & Olsen, C. W. (2014). Building institutional capacities and competencies for systemic learning analytics initiatives. In *Proceedings of the Fourth International Conference on Learning Analytics and Knowledge* (pp. 257–260).

Brown M, McCormack M, Reeves J, Brook DC, Grajek S, Alexander B, Bali M, Bulger S, Dark S, Engelbert N, Gannon K (2020). *2020 educause horizon report teaching and learning edition*.

Brynjolfsson, E., & McAfee, A. (2014). WW Norton & Company. In *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*.

Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics*, *6*(2), 1–9.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. *18*). John Wiley & Sons.

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*(2), 205–235.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*, 319–340.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114–126.

Goldstein, P. J., & Katz, R. N. (2005). *Academic analytics: The uses of management information and technology in higher education* (Vol. *8*). Educause.

Grunhut, J., Wyatt, A. T. M., & Marques, O. (2021). Educating future physicians in artificial intelligence (AI): An integrative review and proposed changes. *Journal of Medical Education and Curricular Development*, *8*, 238212052110368.

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). Association for Computing Machinery.

Lee, J., & See, K. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lewis, J. D., & Weigert, A. (1985). Trust as a social reality. *Social Forces*, *63*(4), 967–985.

Lind, E. A., & Tyler, T. R. (1988). *The social psychology of procedural justice*. Springer Science & Business Media.

McCoy, L. G., Nagaraj, S., Morgado, F., Harish, V., Das, S., & Celi, L. A. (2020). What do medical students actually need to know about artificial intelligence? *NPJ Digital Medicine*, *3*(1), 1–3.

Morewedge, C. K., Yoon, H., Scopelliti, I., Symborski, C. W., Korris, J. H., & Kassam, K. S. (2015). Debiasing decisions: Improved decision making with a single training intervention. *Policy Insights from the Behavioral and Brain Sciences*, *2*(1), 129–140.

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring Teachers' Trust in AI-based educational technology. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 56–66). Association for Computing Machinery.

Nazaretsky, T., Cukurova, M., Ariely, M., & Alexandron, G. (2021). Confirmation bias and trust: Human factors that influence teachers' adoption of AI-based educational technology. In *Companion Proceedings of the Sixteenth European Conference on Technology Enhanced Learning. AI for Blended-Learning:Empowering Teachers in Real Classrooms Workshop, Vol. 3042. CEUR Workshop Proceedings (CEUR-WS.org)*.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220.

OECD. (2019). OECD future of education and skills 2030, conceptual learning framework: Knowledge for 2030. http://www.oecd.org/education/2030-project/teaching-andlearning/learning/knowledge/Knowledge_for_2030_concept_note.pdf

Pardo, A., & Siemens, G. (2014). Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, *45*(3), 438–450.

Piderit, S. K. (2000). Rethinking resistance and recognizing ambivalence: A multidimensional view of attitudes toward an organizational change. *Academy of Management Review*, *25*(4), 783–794.

Powell, A., Watson, J., Staley, P., Patrick, S., Horn, M., Fetzer, L., Hibbard, L., Oglesby, J., & Verma, S. (2015). Blending learning: The evolution of online and face-to-face Education from 2008-2015. Promising practices in blended and online learning series. *International Association for K-12 Online Learning*. https://eric.ed.gov/?id=ED560788

Rienties, B. (2014). Understanding academics' resistance towards (online) student evaluation. *Assessment and Evaluation in Higher Education*, *39*(8), 987–1001.

Sellier, A.-L., Scopelliti, I., & Morewedge, C. K. (2019). Debiasing training improves decision making in the field. *Psychological Science*, *30*(9), 1371–1379.

Tsai, Y.-S., & Gasevic, D. (2017). Learning analytics in higher education—Challenges and policies: A review of eight learning analytics policies. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 233–242).

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131.

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.

Wartman, S. A., & Combs, C. D. (2018). Medical education must move from the information age to the age of artificial intelligence. *Academic Medicine*, *93*(8), 1107–1109.

Wolff, A., Moore, J., Zdráhal, Z., Hlosta, M., & Kuzilek, J. (2016). Data literacy for learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 500–501).

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1–12).

How to cite this article: Nazaretsky, T., Ariely, M., Cukurova, M. & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 00, 1–18. https://doi.org/10.1111/bjet.13232