JOURNAL OF LEARNING ANALYTICS

Volume 10(1), 37–53. https://doi.org/10.18608/jla.2023.7801

The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation

Rianne Conijn[1], Patricia Kahr[2], Chris Snijders[3]

Abstract

Ethical considerations, including transparency, play an important role when using artificial intelligence (AI) in education. Explainable AI has been coined as a solution to provide more insight into the inner workings of AI algorithms. However, carefully designed user studies on how to design explanations for AI in education are still limited. The current study aimed to identify the effect of explanations of an automated essay scoring system on students' trust and motivation. The explanations were designed using a needs-elicitation study with students in combination with guidelines and frameworks of explainable AI. Two types of explanations were tested: full-text global explanations and an accuracy statement. The results showed that both explanations did not have an effect on student trust or motivation compared to no explanations. Interestingly, the grade provided by the system, and especially the difference between the student's self-estimated grade and the system grade, showed a large influence. Hence, it is important to consider the effects of the outcome of the system (here: grade) when considering the effect of explanations of AI in education.

Notes for Practice

- Transparency in algorithms can be provided via explainable artificial intelligence.
- No effect was found of explanations on student trust and motivation in an automated essay scoring system.
- Student trust and motivation was largely influenced by the grade provided by the system.
- There is no one-size-fits-all explanation for different stakeholders and contexts.

Keywords

Explainable artificial intelligence, automated essay-scoring systems, trust, motivation, academic writing

Submitted: 29/11/2022 — **Accepted:** 20/01/2023 — **Published:** 08/03/2023

Corresponding author [1]Email: m.a.conijn@tue.nl Address: Human–Technology Interaction Group & Eindhoven Artificial Intelligence Systems Institute, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. ORCID ID: https://orcid.org/0000-0002-6316-4892

[2]Email: p.k.kahr@tue.nl Address: Human–Technology Interaction Group, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands. ORCID ID: https://orcid.org/0000-0002-1368-8698

[3]Email: c.c.p.snijders@tue.nl Address: Human–Technology Interaction Group, Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands.ORCID ID: https://orcid.org/ 0000-0001-6165-7645

1. Introduction

Artificial Intelligence (AI) algorithms are increasingly used in education; for example, in intelligent tutoring systems or automated grading of essays (Allen et al., 2016). The increasing use of AI, or more generally, the use of learning analytics (LA) in educational institutions, has also led to a rise in ethical considerations in the LA community, including privacy, fairness, equity, and transparency (Ferguson et al., 2016; Cerratto Pargman & McGrath, 2021). This has resulted in the development of several ethical frameworks as well as — to a lesser extent — empirical studies on the implementation of ethical considerations (Cerratto Pargman & McGrath, 2021).

One common consideration in ethical frameworks for LA is the notion of transparency, albeit with different definitions. Slade and Prinsloo (2013) describe transparency as a form of informed consent that involves educational institutions being open about the purposes of their collection and usage of educational data and about how student privacy is ensured. In their follow-up work, Prinsloo and Slade (2018) coin the term "responsible" learning analytics, part of which includes transparency about the criteria and processes. For example, an automated grading system would need to explain how it came up with a specific grade. Sclater's (2016) code of practice also reflects a broader definition of transparency by emphasizing the need to

be open about all aspects of the use of learning analytics, including the processes and algorithms involved in providing analytics to staff and students. It has been argued that a lack of transparency will result in lower trust and higher resistance towards AI (Tzimas & Demetriadis, 2021). Empirical studies have shown that a lack of perceived transparency also reduces teachers' trust in AI (Nazaretsky, Cukurova, et al., 2022; Qin et al., 2020). Trust, in turn, has been shown to affect the user's acceptance of AI technology (Choi et al., 2023).

Transparency about (AI) algorithms in education touches on the more general need for "algorithmic accountability" (Knight et al., 2018). Knight et al. (2018) argue that when the functioning of an algorithm becomes a concern, it is important to be transparent about its inner workings. One criticism of automated systems in education is that simply being told what to do does not constitute learning or teaching, and might even be experienced as disempowering (Ferguson, 2019; Selwyn, 2019). It is important to explain to students why the system made specific choices (Selwyn, 2019). Knight et al. (2018) describe three ways to provide transparency about the inner workings of algorithms to stakeholders: 1) open research publication, 2) open testable systems, and 3) open stakeholder communication. The latter is especially important for students and teachers since they will not benefit from (interacting with) the source code (Knight et al., 2018). In their study, open stakeholder communication was created through co-design of the system with the stakeholders, including extensive discussions with the researchers. In the current study, we examine how we can provide explanations about the inner workings of an AI system to its users in an educational setting as it is being used (as opposed to during the design phase), and explore how these explanations can empower learners.

Providing users with insight into the inner workings of algorithms is not new. In a variety of domains, including self-driving cars and medical decision-making, researchers have used explanations "to open the black box" (Samek & Müller, 2019). This approach has also been referred to as explainable AI (XAI). XAI has several advantages, including improving understanding of the algorithm, improving perceived competence, increasing or calibrating trust, improving acceptance of the outcomes of the algorithm, and enhancing decision-making based upon the outcomes (e.g., Cramer et al., 2008; Ribeiro et al., 2016). However, carefully designed explanations are still in the development stage (Mueller et al., 2019), especially in the field of education. One notable difference between using XAI in education, compared to other domains, is that the explanations should not only increase trust in the system or the system's decision making, but should also empower students to perform better (cf. Ferguson, 2019; Selwyn, 2019). Therefore, the current study aims to identify how explanations for AI in education affect student trust in the algorithm as well as their motivation to continue learning; in this case, we are using an automated essay scoring system.

Automated essay scoring systems are widely known to support writing instruction and assessment (Allen et al., 2016; Hussein et al., 2019). They provide an automated score that can be used to replace or assist teachers in assessing writing quality (Dikli, 2006). Often, the models are trained on a corpus of human-graded essays. Natural language processing is used to identify different features in the written text, followed by statistical modelling (e.g., regression models, see Attali & Burstein, 2004) or machine leaning (e.g., neural networks, see Alikaniotis et al., 2016) to predict the essay score. The systems show high accuracies, with correlations ranging from 0.80 to 0.97 between human and machine ratings (Hussein et al., 2019). Despite this high accuracy, critics have argued that these systems might not measure the critical constructs, could be easily fooled (gaming the system), and lack "the human element" in grading (Dikli, 2006; Hussein et al., 2019). These factors in turn could influence students' trust in the system and their motivation to act upon the feedback. This study determines how explanations via XAI could ameliorate these effects using the following research questions:

RQ1. What is the effect of explanations for automated essay scoring on student trust in the system?

RQ2. What is the effect of explanations for automated essay scoring on student motivation?

2. Background

2.1. Trust in Automation

Trust in automation has been 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. 51). Given the current high accuracies of automated essay scoring systems, a high level of trust would be warranted. However, even in cases where systems show high accuracy, we see that humans tend to have low trust (Prahl & van Swol, 2017). Since trust is a key component in human–computer interaction, researchers have already addressed a variety of factors that affect trust, including transparency or openness about the inner workings of the algorithm. In general, explaining results and recommendations has been shown to increase trust in a system, such as by providing examples (Fenster et al., 2012), causal explanations (Shin, 2021), or visualizations (Yang et al., 2020). Dietvorst et al. (2018) found that explanations increased user trust by giving them more of a sense of control over the system. Understanding technical AI capabilities positively correlated with trust (i.e., Chao et al., 2016; Alan et al., 2014). In contrast, lack of transparency resulted in guessing and attempts at gaming a system (Möhlmann & Zalmanson, 2017).

An additional reason for increasing transparency and making systems comprehensible to users is that trust is highly influenced by user perceptions of the algorithm's accuracy. After seeing a system err, user trust can easily decrease, to the extent that users refuse to rely on a system (cf. Dietvorst et al., 2015; Dzindolet et al., 2003; Manzey et al., 2012). Restoring user trust requires substantial effort, however, which makes preventing loss of trust all the more important. Additionally, research has studied ways to communicate after a system failure, for instance, by using different repair strategies, apologies, promises, justifications, or explanations with different attributions (cf. Kim & Song, 2021; Esterwood & Robert, 2021; Matzat & Snijders, 2012). Apologies seem especially able to repair trust, or at least prevent steep losses (Kim et al., 2004; Snijders et al., 2017; Wang et al., 2016).

2.2. User Motivation and Automation

One often-mentioned benefit of using automation in education is that it can be used to provide students with (additional) feedback in a timelier, personalized way (see e.g., Allen et al., 2016). Research has shown that the use of automated essay scoring, on top of teacher evaluation, can enhance students' writing motivation (Wilson & Czik, 2016). However, when the writing feedback is inaccurate or misunderstood, it can decrease writing motivation (Zumbrunn et al., 2016). Providing an explanation for the feedback could decrease this effect. This is in line with the general idea that feedback is more effective when more information is provided (Hattie, 2008; Wisniewski et al., 2020). Such explanations can be seen as additional since feedback is defined as any information provided to a learner on their performance or understanding (Hattie & Timperley, 2007). Moreover, these explanations are key to enabling students to act upon the system's decision or score (Khosravi et al., 2022) and could be seen as tools to improve self-regulated learning skills, thus aiding students to act in more self-motivated and empowered ways (cf. Ferguson, 2019; Selwyn, 2019). Within intelligent tutoring systems, providing explanations on student knowledge, misconceptions, and difficulties via open learner models can indeed support self-regulated learning (Long & Aleven, 2013).

2.3. Explainable Artificial Intelligence

The goal of XAI is to foster comprehensibility of algorithms, especially for users without expert AI knowledge (Mohseni et al., 2021). XAI aims to create a toolbox of approaches and techniques to support human agents "to understand how, when, and why predictions are made" (Kamath & Liu, 2021, p. 2), for instance by adding (non-technical) explanations, interpretations, and justifications to machine outputs. By definition, AI-based user support is only as good as the understanding and acceptance of users, especially for those who are not tech-savvy (Arrieta et al., 2020). Yet, the format, content, and timing of explanations may vary for different stakeholders and contexts (Clancey & Hoffman, 2021). Accordingly, several researchers stress the importance of incorporating user perspectives in the design process. Eiband et al. (2018) created a stage-based participatory process in which the content (what to explain) and format (how to explain) are iteratively evaluated with users. Recently, a design framework was specifically tailored to XAI in education: the XAI-ED framework (Khosravi et al., 2022). This framework goes beyond focusing on how and what to explain by describing six guiding questions: How are the stakeholders involved? What are the benefits for each group? How is an explanation delivered? What are the models used in education? What needs to be considered for designing the AI (interface) to support the explanation? What are potential pitfalls when providing explanations?

In line with such approaches, we first involved stakeholders (students) in the design of effective explanations for automated essay scoring. We focused on the goal (what are the benefits), the content (how is an explanation delivered), and the format (designing the AI interface). In addition, we employed frameworks and guidelines available from the literature on these three aspects, as described below.

The goal of the explanation could be 1) to justify outcomes; 2) to enhance the capability of identifying system errors; 3) to improve the algorithm, and 4) to learn new knowledge for a specific task at hand (Adadi & Berrada, 2018). A similar guideline from Meuwissen and Bollen (2021) proposes that XAI must aim to answer at least one of four questions: 1) What factors play a role; 2) How does the AI system derive at its conclusion; 3) Why does the AI make a specific decision, and 4) Why not another decision. The goals of the explanation can also directly inform the necessary content. For example, to justify outcomes, or to identify what factors play a role, there may be no need to explain the exact inner workings of the system. Moreover, distinctions have been made regarding the method of explanation (Arrieta et al., 2020; Rawal et al., 2022). Explanations can be textual, and visual; can focus on an, example, counterfactual, or simplification (e.g., using a simplified model); can provide feature relevance, knowledge rules, full system descriptions (global explanations), or provide reasoning for a specific decision made by the system (local explanations).

Some frameworks and guidelines specifically focus on the content of the explanation. For example, Bellotti and Edwards (2001) propose a design framework that stresses the need for intelligibility and accountability for users to make informed decisions. Based on this, they recommend four design principles: 1) inform users about system capabilities and understandings;

2) provide feedback including consequences of actions; 3) enforce identity and action disclosure (e.g., when sharing information), and 4) provide control to the user, which is especially important in cases of conflict. Similarly, Lim and Dey (2010) developed the Intelligibility Toolkit to help developers apply the right form of explanation for context-aware applications. Based on the most common decision model types (rules, decision trees, naïve Bayes, hidden Markov models) and eight explanation types (Inputs, Outputs, What, What If, Why, Why Not, How To, Certainty), the toolkit supports developers with the generation of explanation structures, querying mechanisms for specify questions, simplifying complex explanations, and presenting explanations to end users. Finally, Mueller et al. (2021) created a Self-Explanation Scorecard, allowing developers to formatively evaluate their explanation, based on eight levels, including no explanation, information on input features (e.g., heat maps), example instances (e.g., recommendations), global information on the algorithm, information on the inner workings of the system, information about system failure or limitations, providing comparisons (e.g., similarities or counterfactuals), and offering failure diagnoses.

Other guidelines focus more on the form of the explanation rather than the content. Kulesza et al. (2015) identified two main principles: explainability (explanations need to be iterative, sound, complete, and not overwhelming) and correctability (explanations must always honour user feedback and need to be actionable, reversible, and communicative of incremental changes). Several of these principles have been tested (and supported) in experimental studies. For example, soundness proved to be important, as poor explanations hinder acceptance (Herlocker et al., 2000). It has been argued that completeness is even more important than soundness for increasing trust (Kulesza et al., 2013). Just providing confidence ratings did not improve labelling in activity recognition (Rosenthal & Dey, 2010) and less detailed explanations resulted in lower trust in clinical decision support systems (Bussone et al., 2015). However, more complete explanations are not always better. For example, fuller explanations were shown to have a positive effect on trust, but also led to over-reliance (Bussone et al., 2015). One study found that accuracy statements resulted in higher trust in the system, compared to full explanations (Papenmeier et al., 2019). Ashoori and Weisz (2019) found that stating the model confidence had a small effect on user trust, while other studies did not find that providing an accuracy statement increased confidence (Cramer et al., 2008; Yin et al., 2019). Hence, it is not known how "complete" an explanation needs to be.

2.4. Explainable Artificial Intelligence in Education

The use of XAI in education is still fairly limited, with a few exceptions. Several researchers argue that the use of open learner models can be seen as some form of explainable AI in education (Conati et al., 2018; Khosravi et al., 2022). Open learner models are often used in intelligent tutoring systems (ITS), where they aim to present the learner model, focusing on the student's knowledge, misconceptions, and difficulties. The goal of these open learner models is to encourage self-assessment and metacognitive processes (Bodily et al., 2018; Bull & Kay, 2016). Barria-Pineda and Brusilovksy (2019) adapted open learner models using visualizations in an online educational system. They found that when applying transparent recommendations, students interacted significantly more often with the system. Further, they infer that the explanations led to higher confidence in attempting learning activities. A review by Bodily et al. (2018) has shown that open learner models are usually fairly simple. The models focus on a single type of data, rarely include behavioural data, and use visualizations based on simple aggregations, such as the percentage of quizzes completed. Therefore, the amount of explanation within these systems is often limited; there is not much complicated content to explain.

Recently, more extensive explanations in a variety of educational systems have been explored. Kizilcec (2016) used textual explanations to show how peer-review grading was used to calculate a grade within a MOOC (N = 120). Not surprisingly, it was found that students who received a low grade trusted the system less. However, he also found that explanations about the procedure reduced this effect. Interestingly, he found that providing too much information reduced trust again. Conati et al. (2021) employed explanations for why students received hints in an ITS tool for mathematics (N = 47). They found that most students preferred to be told why the system gave a hint over how the system chose it. In addition, these explanations resulted in a higher intention to use hints, increased perceived helpfulness of the hints, and a higher trust in the system to provide appropriate hints. Ooge et al. (2022) studied the effect of explanations on student trust while they interacted with an e-learning platform that recommended mathematics exercises (N = 37). The explanations were created via a user-centred design process, and included a why statement explaining why the student received the recommendation and a justification of the estimated number of tries needed, as well as a histogram showing the number of tries needed by similar students. Results showed that real explanations (versus placebo explanations or no explanations) significantly increased trust.

Some explanations are tailored to teachers rather than students. For example, Nazaretsky, Arieli, et al. (2022) examined teachers' trust in a so-called AI-grader, which used natural language processing and AI to assess student responses to open-ended questions. The explanations included information on how the decisions were made (including an analytical rubric showing the main content and causal relations that should appear in a correct response) and how the AI can complement the teacher. The interviews (N = 6) showed that these explanations reduced teacher concern and improved trust. Alonso and

Casalino (2019) applied the XAI data-mining tool ExpliClass to facilitate decision-making processes based on predictions of student performance. They included both graphical and textual explanations of the predictions, arguing that these would be suitable for multiple stakeholders — including teachers, tutors, students, and managers — to make informed decisions. However, no experimental results were presented (Alonso & Casalino, 2019).

Initial evidence shows that explanations using open learner models or XAI could enhance trust and motivation in educational technologies. However, empirical evidence on this point is still rather limited, and sometimes based on anecdote or small case studies. Moreover, it is unknown how complete an explanation needs to be. Finally, user perspectives on what to explain and how to explain it are often ignored (Eiband et al., 2018; Khosravi et al., 2022). Therefore, we employed a user-centred approach to design the explanations for AI in education. Specifically, for automated essay scoring systems, we compare full-text explanations and an accuracy statement versus no explanations to examine student trust and motivation.

3. Methods

3.1. Design

Our experiment tests the effect of explanations in automated essay scoring systems on user trust and motivation. Participants were asked to write a short argumentative essay in a controlled online environment and, after completion, to estimate a grade for their own essay. Then an essay scoring system provided a grade, with an explanation. Finally, participants were allowed to adapt their own grade. The study followed a between-subjects design with three conditions: a full-text explanation, an accuracy statement, and a control condition without any explanation. In addition, the explanations differed regarding their level of anthropomorphism (not present versus present), including a voice agent and informal language. This factor is beyond the scope of the current paper, and proved to have no substantial effect on the results, hence has been left out of the analysis. Trust and motivation were the dependent variables, both measured using objective behavioural metrics and subjective questions. The study was approved by the Ethical Review Board of Eindhoven University of Technology.

3.2. Participants

The sample size was calculated through an a priori power analysis in G*Power 3.1. Based on previous studies, we expected a small to moderate effect size (f = 0.3). To reach a power of 90%, the sample size required is at least 144 participants. We recruited via the Participant Database of Eindhoven University of Technology. In total, 152 university students (71 females, 81 males) participated in the study and received compensation for participating — €5 or study credits (their choice). Two participants were excluded from the data analysis since they wrote fewer than 100 words (at least 150 words were required) or spent less than seven minutes on writing the essay.

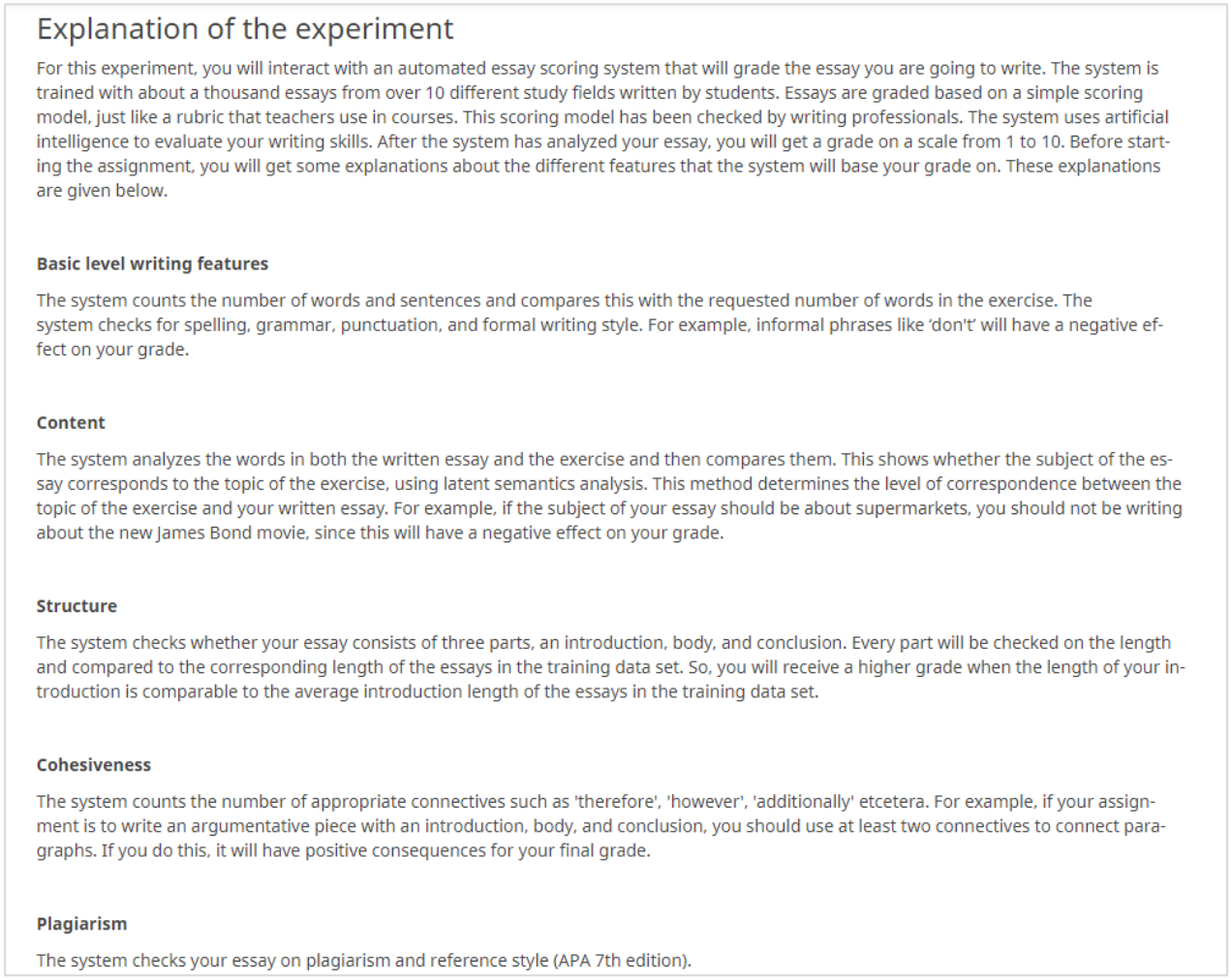
3.3. Materials

Participants were informed that they were interacting with an actual AI essay grading system. However, the underlying system only consisted of some simple rules integrated within LimeSurvey, rather than being a full AI system (Wizard-of-Oz AI). To have more control over the grades, we chose not to use a full AI grading system. To reduce the variance in the grades, only three grades were given: 4.2, 6.2, or 8.2 (on a scale of 1 to 10). Participants received 8.2 if they included all pre-coded key topic words and met the word limit, 6.2 if they met the word limit but did not include all topic words, and 4.2 otherwise.

We employed a user-centred approach to design the explanations by focusing on the goals (benefits), content, and form of the explanations (cf. Khosravi et al., 2022). We focused on one group of stakeholders: students. The main goal of the explanations was defined as being able to learn new knowledge for a specific task at hand (Adadi & Berrada, 2018). In this, explanations serve as additional feedback for students. We made sure to explain what factors play a role in the automated writing evaluation (Meuwissen & Bollen, 2021). The stakeholder needs were identified using two focus groups (N = 11) of university students. The needs-elicitation study showed that students were less interested in explanations of surface features (such as spelling or grammar), as they were already familiar with automated grading of these features by word-processing software, but were more interested in higher-level features such as content, structure, and cohesiveness. In addition, the focus group students argued that the explanations should be relatable.

The form of the explanations was different for the two explanation conditions, where the full-text explanation included a complete and sound explanation (cf. Kulesza et al., 2015) and the accuracy statement only included a confidence score. The confidence score was based on previous studies, which found that the correlation between human graders and the automated essay scoring systems was about 90% (see Section 2.1). Hence, the system was said to be accurate 90% of the time.

The content of the full-text explanation was geared to what factors are important in essay grading and how the algorithm was trained (see Figure 1). We emphasized the higher-level features, as inspired by the focus groups, including content, structure, cohesiveness, and plagiarism, and combined the surface level features into one dimension: “basic level writing features.” To make the explanations relatable, one option would be to provide explanations tailored to the specific student’s



Explanation of the experiment

For this experiment, you will interact with an automated essay scoring system that will grade the essay you are going to write. The system is trained with about a thousand essays from over 10 different study fields written by students. Essays are graded based on a simple scoring model, just like a rubric that teachers use in courses. This scoring model has been checked by writing professionals. The system uses artificial intelligence to evaluate your writing skills. After the system has analyzed your essay, you will get a grade on a scale from 1 to 10. Before starting the assignment, you will get some explanations about the different features that the system will base your grade on. These explanations are given below.

Basic level writing features

The system counts the number of words and sentences and compares this with the requested number of words in the exercise. The system checks for spelling, grammar, punctuation, and formal writing style. For example, informal phrases like 'don't' will have a negative effect on your grade.

Content

The system analyzes the words in both the written essay and the exercise and then compares them. This shows whether the subject of the essay corresponds to the topic of the exercise, using latent semantics analysis. This method determines the level of correspondence between the topic of the exercise and your written essay. For example, if the subject of your essay should be about supermarkets, you should not be writing about the new James Bond movie, since this will have a negative effect on your grade.

Structure

The system checks whether your essay consists of three parts, an introduction, body, and conclusion. Every part will be checked on the length and compared to the corresponding length of the essays in the training data set. So, you will receive a higher grade when the length of your introduction is comparable to the average introduction length of the essays in the training data set.

Cohesiveness

The system counts the number of appropriate connectives such as 'therefore', 'however', 'additionally' etcetera. For example, if your assignment is to write an argumentative piece with an introduction, body, and conclusion, you should use at least two connectives to connect paragraphs. If you do this, it will have positive consequences for your final grade.

Plagiarism

The system checks your essay on plagiarism and reference style (APA 7th edition).

Figure 1. Full explanations of the automated essay scoring system

text (i.e., local explanations). However, since we used a Wizard-of-Oz automated essay grading system, it was not possible to provide local explanations; hence, we used global explanations, in combination with a more general example for each feature to show how this could affect their grade. Combined, the full-text explanation including information on Inputs, Outputs, What, and What If (Lim & Dey, 2010) to reach level 4 on the Self-Explanation Scorecard (Mueller et al., 2021). The explanations were limited in that they did not provide any control to the user (cf. Bellotti & Edwards, 2001), nor any counterfactual explanations or information on system limitations (cf. Lim & Dey, 2010; Mueller et al., 2021).

3.4. Measurement

The two dependent variables, trust and motivation, were both measured in more than one way. **Situational trust**, or "context-dependent trust," was measured using the Trust in Automation (TiA) scale (Jian et al., 2010). This questionnaire consists of 12 items (e.g., "I am confident in the system") ranging from 1 (not at all) to 7 (extremely). The scale was reliable ($\alpha = .88$). In addition, trust was measured using the Weight of Advice theorem (WoA), which reflects the degree to which the participants adapt their own judgment of their estimated essay grade towards the grade given by the system (Hütter & Ache, 2016). Here, changing the grade towards the estimate from the system is seen as a sign of trust (Papenmeier et al., 2019).

$$Objective\ trust = \frac{grade_{second\ estimate} - grade_{first\ estimate}}{grade_{system} - grade_{first\ estimate}}$$

Motivation was assessed on the interest/enjoyment scale from the Intrinsic Motivation Inventory (McAuley et al., 1989). The interest subscale consists of seven questions (e.g., "I found writing the essay very interesting"), ranging from 1 (strongly

disagree) to 7 (strongly agree). The scale showed high reliability (α =.91). In addition to this measure, participants were also asked if they wanted to continue working on their essays after having received the grade from the system.

Several other measures were included that might influence the dependent variables, including dispositional trust, need for cognition, system grade, and difference in grade. **Dispositional trust** was measured with the Propensity to Trust Technology (PTT) scale (Jessup et al., 2019). The scale consists of six items (e.g., "Generally, I trust technology"), which were measured on a scale ranging from 1 (strongly disagree) to 7 (strongly agree). The Cronbach's alpha was .70, indicating that the scale is acceptable. **Need for cognition**, or one's tendency to engage in effortful cognitive activity, was measured by the Need for Cognition scale (Lins de Holanda Coelho et al., 2020). The scale consists of six items (e.g., "I like to have the responsibility of handling a situation that requires a lot of thinking"), measured on a scale ranging from 1 (extremely uncharacteristic of me) to 7 (extremely characteristic of me). The scale was reliable (α = .83). Two measurements relate to the **grade** provided by the system, as low grades could negatively influence trust and/or motivation. First of all, we added the system grade itself as a covariate. Second, we included the difference between the system grade and the student's self-estimated grade.

The **quality of the explanations** was measured using the explanation quality questionnaire, adapted from Conati et al. (2021). Two questions were excluded as they related to the navigation to the explanation (in our study, the explanations were always shown, so no navigation was needed). In total, this left three items related to the usefulness of the explanations, three related to their intrusiveness, and one related to their usability. A confirmatory factor analysis with oblique rotation ("promax") showed that these scales could not be recreated in our dataset; therefore, we will report on the quality using individual items from this item set, rather than on a scale.

Furthermore, all essays were **graded** by two independent raters, where 20% of the essays were graded twice. The inter-rater reliability was good, with the intraclass correlation (ICC) = 0.757. Both raters graded the essays using a rubric, following the same features as the automated essay grading system: content, structure, cohesiveness, plagiarism, and basic level writing features. Finally, we measured the **total time spent in explanations** to get an indication of how much effort the participants put into reading the explanations.

3.5. Procedure

After signing the informed consent, students were led to a cubicle to start the experiment, which began with a short questionnaire on demographics and propensity to trust. Participants were randomly divided into one of the explanation conditions. In the control condition, participants were only informed that they were graded by an automated essay grading system, without further explanation. In the explanation conditions, participants received a full-text explanation or an accuracy statement of the system. Thereafter, participants were asked to write an argumentative essay about blocking snack bars near primary schools to improve children's health. They received a short news article as a source and asked to write 150–250 words within 20 minutes. After writing the essay, participants filled in the motivation questionnaire and asked to estimate their own grade. Thereafter, participants received the system grade, with or without explanation, depending on the condition.

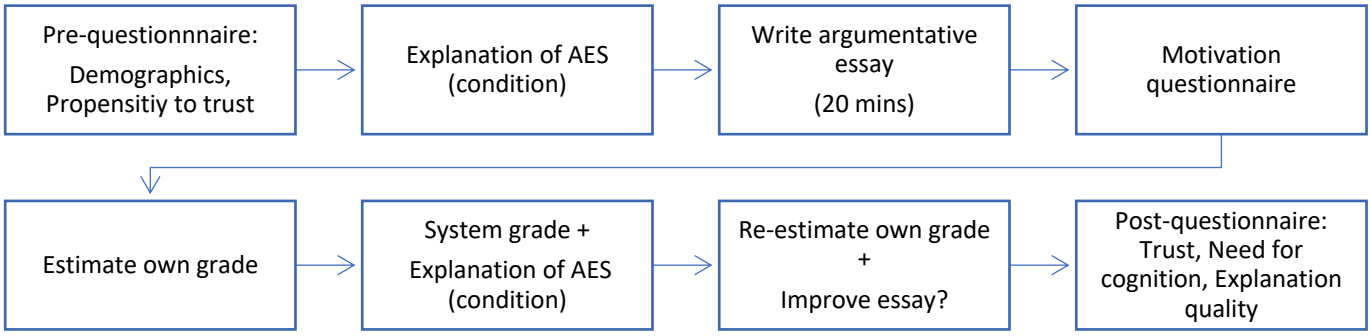


Figure 2: Overview of the experimental procedure

After the system grade, students were asked if they wanted to change their initial grade or if they wanted to improve their essay. If they wanted to improve their essay, they were thanked for their enthusiasm, but they did not actually need to work more on the essay. Thereafter, participants filled out several surveys related to their trust in the automated essay scoring system, their need for cognition, and explanation quality (for those in the full explanation condition). Finally, participants were debriefed and thanked for participation. The complete experiment took 30 minutes in total. An overview of the experimental procedure can be found in Figure 2.

4. Results

4.1. Descriptive Statistics

The full explanations were shown to be of reasonable quality (see Figure 3). On average, students liked the explanations and said that they were not very distracting or confusing. However, the full explanations were also considered to be quite overwhelming and not that clear. Students spent, on average, 2 minutes and 40 seconds on the full explanations (SD = 53 s), compared to 61 seconds (SD = 26 s) for the accuracy statement explanation.

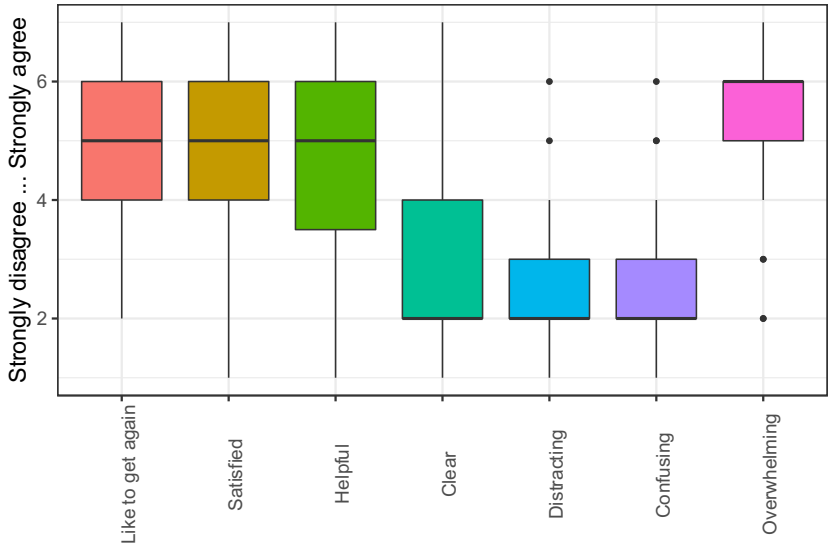


Figure 3. Boxplots of the quality of the full explanations (n = 51)

Overall, the students had a high initial trust (M = 5.64, SD = 0.61) and a high need for cognition (M = 5.33, SD = 0.88). On average, they wrote between 117 and 389 words (M = 207, SD = 42), and took 7 to 20 minutes (M = 15.2, SD = 3.6). The students rated their own essays, on average, at 6.5 (SD = 0.98) out of 10; the human graders rated them at 7.5 (SD = 0.69) out of 10. Table 1 shows the descriptive statistics for both trust measurements and both motivation measurements across the three conditions. It can be seen that the trust in the system is average (around 4.0 on a scale from 1 to 7). Most participants (54–69%) changed their grade towards the grade of the system. Their motivation proved to be below average, and only 20–25% were motivated to continue their writing. Some evidence shows that a full explanation works better than either the accuracy statement or the control for all four measures, except for the motivation to continue. The differences are small, though. We now consider these differences in more detail and in a multivariate analysis.

Table 1. Descriptive Statistics of Trust and Motivation Across Conditions

	N	Trust (subjective) M (SD)	Adapted grade towards system (%)	Motivation (subjective) M (SD)	Motivation to continue (%)
Control	48	3.98 (0.95)	54%	3.68 (1.07)	25%
Accuracy statement	51	3.87 (0.82)	63%	3.49 (1.16)	22%
Full explanation	51	4.24 (1.01)	69%	3.76 (1.02)	20%

4.2. Effects on Trust

We first considered the effect of explanations on students' subjective trust of the system and found that the explanation factor could not predict trust ($F(2,147) = 2.14$, $p = 0.12$). Both the accuracy statement ($p = 0.82$) as well as the full explanation ($p = 0.15$) did not lead to higher trust in the automated essay scoring system compared to the system without explanations. Thereafter, we determined the effect of several potential predictor variables on trust (see Table 2). Again, no effects were found for explanations on trust, not even when controlling for the time spent reading the explanation. The difference between a student's own estimated grade, compared to the grade given by the system, played an important role in their trust, but appeared to have a non-linear effect. Model comparisons with up to third-degree polynomials showed that a quadratic function has the best fit. When the system predicts a lower grade than the user, the trust is lower, compared to when the system predicts

a higher grade than the user. However, this effect decreases when the system produces a much higher grade than the user, which makes intuitive sense (for a graphical overview of the effect, see the left part of Figure 4). While the system grade had a significant effect, the actual grade as scored by the human grader did not affect trust. Finally, for the personal characteristics, propensity to trust was shown to have a significant effect on subjective trust (B = 0.25, SE = 0.11, p = 0.020). Participants with a higher propensity to trust showed more trust in the automated essay grading system. Regarding the need for cognition, one could argue that students with a higher need for cognition would benefit more from the full explanations. However, this interaction effect turned out to be non-significant (p = 0.26).

Thereafter, we determined the effect of explanations on the size of students' adaptation of the grade. More than one-third (55 students) did not adapt their initial estimate at all, and one student even adapted the grade away from the grade given by the system. Because of this highly skewed distribution, we decided to treat this measure as a binary variable, where students are defined as trusting the system if they adapted their grade in the direction of the system's grade. Interestingly, no relation was found between subjective trust and student willingness to act upon the system (B = 0.12, SE = 0.16, p = 0.44). To determine the effect of explanation on willingness to change the grade towards the advice, we used logistic regression. Although more students adapted their grade when they received an accuracy statement, and even more when they received a full explanation, these differences were found to be non-significant. Thereafter, we included the same predictor variables as for subjective trust, except for the difference in grade (see Table 2). First, even when controlling for time spent on the explanations, no effects were found for the explanation on the likelihood of changing the grade towards the system. Higher system grades correlated with a higher likelihood of changing the initial grade (p = 0.03).

Figure 5 (left) shows the distribution of difference in grades for students who did and did not adapt their grade. Here we see a binomial distribution, where the absolute difference in grade appeared to be more important than the relative difference. Hence, we divided the difference into two predictors: binary (whether the system grade was lower than the self-estimated grade) and absolute, again modelled with a quadratic function. A larger difference between the system grade and the participant's grade resulted in a higher likelihood to change the grade towards the system, but again this effect was reduced when the difference became much larger. Again, the human grades had no effect on the likelihood of changing the grade. Finally, propensity to trust had a positive effect on trust (p = 0.02). Need for cognition, as well as the interaction effects with the conditions, did not have a significant effect on trust. Overall, the model was able to correctly classify whether a student changed the grade in 81% of the cases. The model was slightly more off for the cases where the participant did not change the grade compared to when they did (precision = 0.83, recall = 0.86).

4.3. Effects on Motivation

We now consider the effect of explanations on students' subjectively measured motivation or interest in continuing the writing assignment. The explanation factor could not predict student motivation ($F(2,147) = 0.83$, $p = 0.43$). Both the accuracy statement ($p = 0.41$) as well as the full explanation ($p = 0.68$) did not lead to higher motivation in the assignment. Table 2 shows a regression analysis that includes the same predictor variables as for trust (except for the propensity to trust). Interestingly, now an effect was found for the explanation, but only for the accuracy statement. The accuracy statement led to higher interest in the assignment, compared to no explanation (B = 2.85, SE = 1.29, $p = 0.03$). This effect did not change with the amount of time spent on reading the explanations. The effect of the accuracy statement was reduced for people with a higher need for cognition, which is consistent with the idea that those with a high need for cognition find just an accuracy statement about as useful as no explanation at all. As with trust, the difference between a student's own estimated grade, compared to the grade given by the system, played an important role in their motivation. Again, model comparisons showed that a quadratic function of difference in grade proved to have a better fit, compared to a linear function. When the system predicts a lower grade than the user, the motivation is higher, compared to when the system predicts a higher grade than the user. However, this effect reduces when the system produces a much higher grade than the user (see Figure 4, right). The human grades did not have an effect on motivation.

We then considered the effect of explanations on students' willingness to continue writing and found a positive relation with subjective motivation (B = 0.67, SE = 0.21, p = 0.002). A logistic regression showed that neither the accuracy statement nor the full explanation led to a higher likelihood to continue writing compared to no explanation at all. In a similar analysis with all other predictors included we found that the system grade, as well as need for cognition, had no effect on the motivation to continue writing, even when controlling for the time spent on explanation. The difference in grade had a negative effect: when the system showed a lower grade than the user's estimate, the participants were somewhat more likely to continue writing (see Figure 5). The human grade as well, as need for cognition, did not have an effect on students' willingness to continue. Given the skewed distribution of students who wanted to continue, the model showed many false negatives, where people were predicted not to continue when they actually would (precision = 0.69, recall = 0.33).

Table 2. Regression Models on Trust and Motivation

	Regression		Logistic regression	
	Subjective trust	Subjective motivation	Changed grade	Continue writing
Explanation type				
No explanation	-	-	-	-
Accuracy statement	-0.21	**2.85***	3.37	4.78
	(1.04)	(1.29)	(3.97)	(3.98)
Full explanation	1.45	1.96	4.06	7.10
	(1.10)	(1.35)	(3.98)	(4.31)
Time spent reading explanation	-0.002	-0.007	0.0008	-0.002
	(0.005)	(0.006)	(0.02)	(0.02)
ExplanationType* Time spent reading explanation				
No explanation * Time spent	-	-	-	-
Accuracy statement * Time spent	0.001	0.007	-0.008	0.016
	(0.007)	(0.008)	(0.02)	(0.02)
Full explanation * Time spent	0.001	0.006	0.001	0.006
	(0.006)	(0.007)	(0.02)	(0.02)
AES grade	-0.11	**0.43*****	0.87*	-0.005
	(0.08)	(0.10)	(0.38)	(0.29)
Grade difference (AES grade – estimated grade)	**5.47*****	**-9.65*****	-	**-14.59***
	(1.54)	(1.90)		**(5.93)**
Grade difference (AES grade – estimated grade)^2	**-4.47*****	**2.80***	-	-
	(0.87)	(1.07)		
System grade lower	-	-	0.93	-
			(0.62)	
Absolute grade difference	-	-	**34.65*****	-
			(6.75)	
Absolute grade difference^2	-	-	**-16.80*****	-
			(4.20)	
Human grade	-0.06	-0.10	-0.26	
	(0.005)	(0.13)	(0.34)	
Propensity to trust	**0.25***	-	0.86*	-
	(0.11)		**(0.38)**	
Need for Cognition	0.07	**0.42****	-0.32	0.86
	(0.13)	(0.16)	(0.44)	(0.53)
ExplanationType* Need for Cognition				
No explanation * N4C	-	-	-	-
Accuracy statement * N4C	0.04	**-0.66****	-0.52	-1.14
	(0.19)	(0.23)	(0.64)	(0.72)
Full explanation * N4C	-0.21	-0.38	-0.68	-1.49
	(0.19)	(0.23)	(0.66)	(0.72)
(Intercept)	3.940	-0.25	-5.79	-4.74
	(1.21)	(1.30)	(3.97)	(3.88)
N	150	150	149	150
R^2	0.27	0.17	-	-
Pseudo R^2 (Nagelkerke)	-	-	0.46	0.26
Accuracy			0.81	0.82

Note: Unstandardized coefficient estimates are presented with SE in parentheses; significant effects are presented in bold; *p < 0.05, **p < 0.01, ***p < 0.001.

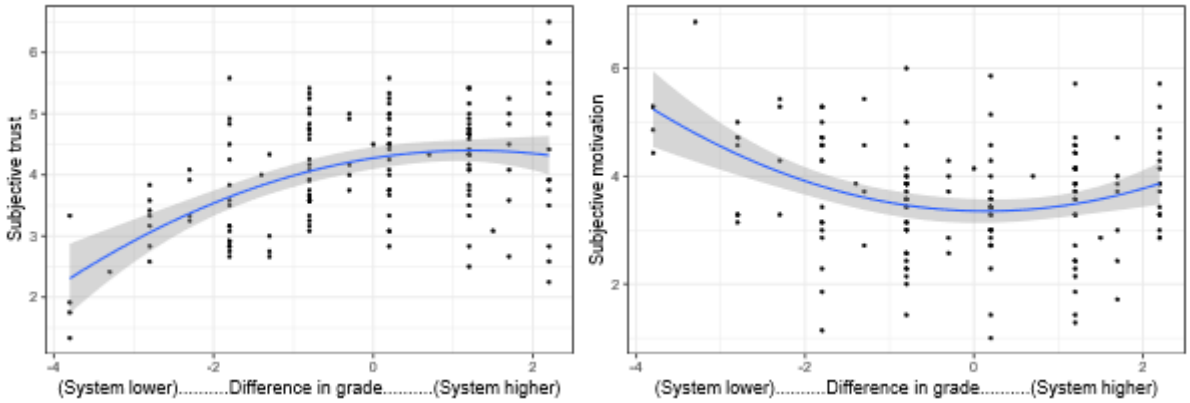


Figure 4. Effect of difference in grade between system and user on subjective trust (left) and motivation (right), both fitted with a quadratic function

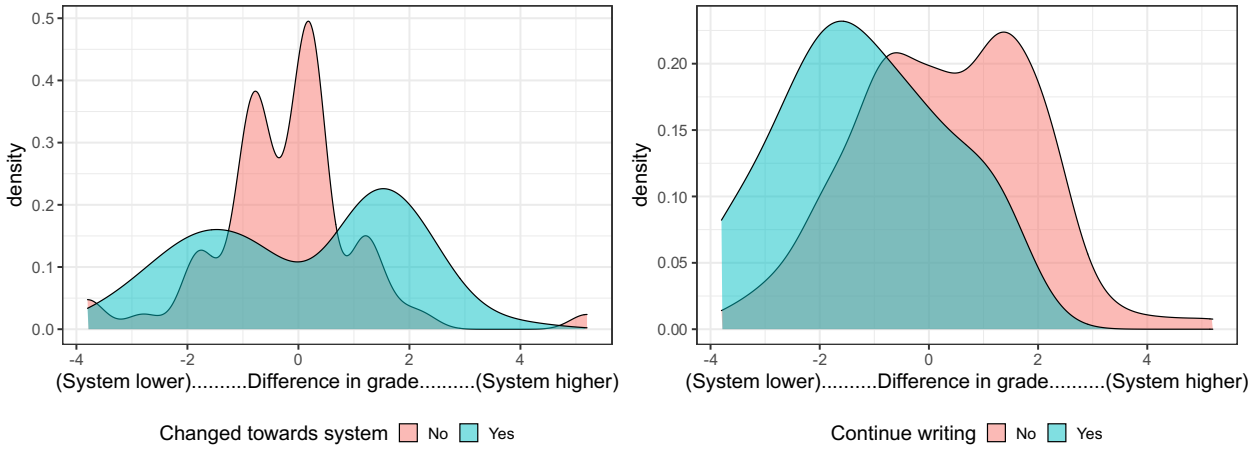


Figure 5. Effect of difference in grade between system and user on grade change towards the system (trust; left) and continuing writing (motivation; right)

5. Discussion and Conclusion

Ethical considerations, including transparency, are increasingly discussed in the learning analytics community. However, the number of empirical user studies examining ethical considerations, such as transparency in AI, is relatively low and often focused on LA in general rather than specific LA systems (Cerratto Pargman & McGrath, 2021). Therefore, we studied transparency in a specific LA system: an automated essay scoring system. Here, transparency was created via an explanation, where students received an automated grade on their own essay, and received either no explanation, an accuracy statement, or a full-text explanation on how the system arrived at its grade. We analyzed how the use of explanations influenced student trust and motivation.

In our experiment, neither an accuracy statement nor a full explanation led to higher trust (neither subjective trust nor willingness to change the grade towards the system), compared to the control condition. This contradicts previous literature, which showed that explanations do result in higher trust in an intelligent tutoring system (Conati et al., 2021; Ooge et al., 2022). This could be because the context was different (intelligent tutoring system vs. automated essay scoring), or because the appropriate format, content, and timing of explanations may differ across contexts (Clancey & Hoffman, 2021). One of the major differences compared to previous research was that Conati et al. (2021) included local explanations, tailored to the specific user, while we used global explanations (the same for all users). In other domains, it has been argued that lay users find local explanations more relatable than global ones (Adadi & Berrada, 2018). Moreover, previous research suggests that students' need for cognition might impact how well they read the explanations and hence their effect (Conati et al., 2021). However, no effect was found for the time spent on reading the explanations, the need for cognition, or the interaction effect on trust. Hence, it does not seem to be the case that students with a higher need for cognition benefit more from the explanations, which might again be due to differences in context.

Other aspects were found to be related to students' trust and their willingness to adapt the grade towards the system. Specifically, initial trust was found to have a positive effect, which corroborates previous work (Campagna et al., 2022; Jessup et al., 2019). Finally, the difference between the estimated grade and the system grade had a large effect on trust, although the pattern differed for different measures of trust. Subjective trust was lower when the system provided a lower grade than the student. For changing the grade towards the system, the direction seemed to matter less than the absolute difference. Students adapted their grade more often when the difference between the system's grade and their own was larger. The effect of grade on trust has been previously established, where lower grades lead to lower trust (e.g., Kizilcec, 2016). Interestingly, this effect was found only for the grade reported by the system. The actual quality of the essay, as scored by human graders (not provided to the participants), did not relate to student trust or motivation. Hence, it could be argued that this effect is mostly due to the displayed grade, rather than the actual essay quality. The effect of the system grade can be explained by the fact that the grading was done on a detailed scale, providing grades up to one decimal point (4.2, 6.2, 8.2). Hence, if the student receives a grade that is very close to the system grade (e.g., 6.0 versus 6.2), this might not even be perceived as a real difference, especially in the Dutch system where grades are often rounded. This proved to be an issue with the measurement of trust via weight on advice (Hütter & Ache, 2016), since this is also affected by the difference between the two grades. For future work, it might be useful to determine the automated essay score based on the user-provided score to be able to identify whether the students adopted the advice.

In addition to trust, we measured the effect of explanations on motivation. Motivation was not explicitly measured via a questionnaire before in the context of AI in education as far as we know, but Conati et al. (2021) showed that explanations of hints resulted in a higher likelihood to use the hints. We did not replicate these findings: no effect of explanations (either full-text explanation or accuracy statement) was found for either subjective motivation or willingness to continue writing. There was only a small positive effect of the accuracy statement (as compared to no explanation) on subjective motivation. Again, these effects can be explained by the different contexts and design of the explanations. In addition, willingness to continue writing also requires a greater effort than using a hint and, hence, might require more effective explanations. In general, student willingness to continue writing was relatively low, which might have resulted in a floor effect. Other factors also had an effect on motivation. Specifically, need for cognition positively influenced motivation, which is in line with the definition of need for cognition: the tendency to engage in and enjoy effortful cognitive activities (Cacioppo et al., 1996). In addition, the difference in grade was related to motivation, but differently than its relation to trust. With a large (absolute) difference between the system grade and the user grade, the subjective motivation is higher. However, subjective motivation was measured before the grade was provided, so no causal relationship can be established. For willingness to continue writing, when the system showed a lower grade than the user's estimate, participants were somewhat more likely to continue writing. This is consistent with students who received lower grades seeing more room for improvement; with a higher grade, there is less need to improve.

5.1. Limitations and Future Work

One of the main limitations is that although the explanations were grounded in a pilot study and consistent with guidelines suggested in the literature, they can certainly be improved upon. For instance, since the automated essay scoring system used in the experiment was a Wizard of Oz system, it was not possible to provide local explanations, specifically tailored to a particular student or essay. In addition, the explanations did not include counterfactuals (e.g., "Your grade would have been higher if …"; cf. Lim & Dey, 2010). Moreover, no information was provided on the system's limitations (Mueller et al., 2021). Finally, the explanations were not interactive, hence, they did not provide any control to the user (cf. Bellotti & Edwards, 2001). Future work should examine additional aspects of explanations to identify which content and forms work in an educational setting. For example, although counterfactuals are becoming more and more popular in XAI (Warren et al., 2022), it might result in undesirable behaviour when used in education, such as gaming the system (cf. Hussein et al., 2019). Other potential pitfalls exist when providing explanations, as exemplified in the XAI-ED framework (Khosravi et al., 2022).

Another limitation is that the current study was focused on one specific type of stakeholder: students. As Clancey and Hoffman (2021) argue, the required format, content, and timing of explanations may differ across different stakeholders (Clancey & Hoffman, 2021). Within their XAI-ED framework, Khosravi et al. (2022) also discuss the need to distinguish between the benefits for different stakeholders. One important stakeholder missing in the current research is teachers, who may use the automated writing evaluation to supplement their own grading. For teachers, effects such as the grade provided by the system potentially have different effects. Therefore, future work should determine the needs of different stakeholders for the design of explanations as well as compare the effects of explanations across different stakeholder groups.

Further, the relatively low scores for motivation may show that students did not take the experiment seriously, which might have lowered the effects of the explanations. However, students did spend a reasonable time on the essay as well as the explanations. In addition, the students wrote essays of medium to high quality, as seen in the human grades. Therefore, the students obviously put reasonable effort in the experiment.

Finally, in the current set-up, students only interacted with the algorithm once. That is, they only wrote one essay and received a single grade. This single interaction might not be enough to generate a clear perceived system accuracy, which is an important factor in user trust. Future work could include multiple interactions with the system to get more reliable data regarding student trust with a single decision and with the system as a whole. This could also be used to identify how quickly trust decreases and can be rebuilt after the system made an error (cf., Dietvorst et al., 2015; Dzindolet et al., 2003).

5.2. Implications and Conclusions

This was the first study to examine the use of explanations in an AI automated essay scoring system. The study showed that the use of full-text global explanations or simple accuracy statements did not lead to higher trust or motivation, compared to no explanations, in this specific context. The grade provided by the system — and specifically the difference between the student's self-estimated grade and system grade — were related to trust and motivation.

Several implications follow from our study. First, it is important to consider the effects of the system outcome (i.e., grade) on explanations in AI in education. Explanations might, for example, be more important when there is a large discrepancy between the expected grade and received grade. Interestingly, a lower system grade had a positive influence on student willingness to continue writing, suggesting that a system might be programmed to be stricter, thus motivating students to continue since they perceive more room for improvement. Second, subjective trust in the system was not found to relate directly to students adapting their grades toward the system's suggestions. Hence, to identify if higher trust also leads to acting on the outcomes, it is important not to focus only on subjective measurements. Finally, this study points towards the idea that there is no one-size-fits-all explanation for AI in education. The effective design of explanations might differ based on a wide variety of factors, including system outcome, type of explanation, context (e.g., the purpose of the AI), and stakeholder. Explanations should not require too much effort nor be overwhelming (cf. Kulesza et al., 2015). Accordingly, more user-centred design, combined with empirical evidence, is needed to determine the effectiveness of explanations, and whether specific decisions on content or form could be generalizable across stakeholders and/or contexts. This would eventually lead to higher transparency, which is of key importance since AI is being increasingly used in education.

Declaration of Conflicting Interest

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors declared no financial support for the research, authorship, and/or publication of this article.

Acknowledgments

The authors would like to thank Julie van der Hijde, Kari Luijt, and Robert Prins for their aid in the design of the explanations and data collection of the experimental study. The authors would like to thank Teis Arets, Suzanne Blokland, Kaia de Groot, and Farida Mesqi for conducting the needs-elicitation study.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alan, A., Costanza, E., Fischer, J., Ramchurn, S. D., Rodden, T., & Jennings, N. R. (2014). A field study of human–agent interaction for electricity tariff switching. *Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems* (AAMAS 2014), 5–9 May 2014, Paris, France (pp. 965–972).

Alikaniotis, D., Yannakoudakis, H., & Rei, M. (2016). Automatic text scoring using neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (ACL 2016), 7–12 August 2016, Berlin, Germany (Vol. 1: Long papers, pp. 715–725). Association for Computational Linguistics. https://doi.org/10.18653/V1/P16-1068

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*, 2nd ed. (pp. 316–329). Guildford Press. https://eric.ed.gov/?id=ED586512

Alonso, J. M., & Casalino, G. (2019). Explainable artificial intelligence for human-centric data analysis in virtual learning environments. In D. Burgos, M. Cimitile, P. Ducange, R. Pecori, P. Picerno, P. Raviolo, & C. M. Stracke (Eds.), *Higher education learning methodologies and technologies online* (pp. 125–138). Springer. https://doi.org/10.1007/978-3-030-31284-8_10

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Bejamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. https://doi.org/10.48550/arXiv.1912.02675

Attali, Y., & Burstein, J. (2004). Automated essay scoring with e-rater® V.2.0. *ETS Research Report Series*, *2004*(2). https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Barria-Pineda, J., & Brusilovsky, P. (2019). Making educational recommendations transparent through a fine-grained open learner model. IUI Workshops '19, 20 March 2019, Los Angeles, CA, USA. https://ceur-ws.org/Vol-2327/IUI19WS-IUIATEC-6.pdf

Bellotti, V., & Edwards, K. (2001). Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction*, *16*(2–4), 193–212. https://doi.org/10.1207/S15327051HCI16234_05

Bodily, R., Kay, J., Aleven, V., Jivet, I., Davis, D., Xhakaj, F., & Verbert, K. (2018). Open learner models and learning analytics dashboards: A systematic review. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge* (LAK '18), 5–9 March 2018, Sydney, NSW, Australia (pp. 41–50). ACM Press. https://doi.org/10.1145/3170358.3170409

Bull, S., & Kay, J. (2016). SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, *26*, 293–331. https://doi.org/10.1007/s40593-015-0090-8

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. *Proceedings of the 2015 IEEE International Conference on Healthcare Informatics* (ICHI 2015), 21–23 October 2015, Dallas, TX, USA (pp. 160–169). IEEE Computer Society. https://doi.org/10.1109/ICHI.2015.26

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253. https://doi.org/10.1037/0033-2909.119.2.197

Campagna, R. L., Mislin, A. A., Dirks, K. T., & Elfenbein, H. A. (2022). The (mostly) robust influence of initial trustworthiness beliefs on subsequent behaviors and perceptions. *Human Relations*, *75*(7), 1383–1411. https://doi.org/10.1177/00187267211002905

Cerratto Pargman, T. C., & McGrath, C. (2021). Mapping the ethics of learning analytics in higher education: A systematic literature review of empirical research. *Journal of Learning Analytics*, *8*(2), 123–139. https://doi.org/10.18608/jla.2021.1

Chao, C.-Y., Chang, T.-C., Wu, H.-C., Lin, Y.-S., & Chen, P.-C. (2016). The interrelationship between intelligent agents' characteristics and users' intention in a search engine by making beliefs and perceived risks mediators. *Computers in Human Behavior*, *64*, 117–125. https://doi.org/10.1016/j.chb.2016.06.031

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

Clancey, W. J., & Hoffman, R. R. (2021). Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. *Applied AI Letters*, *2*(4). https://doi.org/10.1002/ail2.53

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503. https://doi.org/10.1016/j.artint.2021.103503

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (*2018).* AI in education needs interpretable machine learning: Lessons from open learner modelling. *Proceedings of the 2018 ICML Workshop on Human Interpretability in Machine Learning* (WHI 2018), 14 July 2018, Stockholm, Sweden (pp. 21–27). https://doi.org/10.48550/arXiv.1807.00154

Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, *18*, 455–496. https://doi.org/10.1007/s11257-008-9051-3

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. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, *5*(1). http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1640

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human–Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (IUI '18), 7–11 March 2018, Tokyo, Japan (pp. 211–223). https://doi.org/10.1145/3172944.3172961

Esterwood, C., & Robert, L. J. (2021, August 12). Do you still trust me? Human–robot trust repair strategies. *Proceedings of the 30th IEEE International Conference on Robot and Human Interactive Communication* (RO-MAN 2021), 8–12 August 2021, Virtual. IEEE Computer Society.

Fenster, M., Zuckerman, I., & Kraus, S. (2012). Guiding user choice during discussion by silence, examples and justifications. *Frontiers in Artificial Intelligence and Applications*, *242*, 330–335. https://doi.org/10.3233/978-1-61499-098-7-330

Ferguson, R. (2019). Ethical challenges for learning analytics. *Journal of Learning Analytics*, *6*(3), 25–30. https://doi.org/10.18608/jla.2019.63.5

Ferguson, R., Hoel, T., Scheffel, M., & Drachsler, H. (2016). Guest editorial: Ethics and privacy in learning analytics. *Journal of Learning Analytics*, *3*(1), 5–15. https://doi.org/10.18608/jla.2016.31.2

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

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

Herlocker, J. L., Konstan, J. A., & Riedl, J. (2000). Explaining collaborative filtering recommendations. *Proceedings of the 2000 Conference on Computer Supported Cooperative Work* (CSCW '00), 2–6 December 2000, Philadelphia, PA, USA (pp. 241–250). ACM Press. https://doi.org/10.1145/358916.358995

Hussein, M. A., Hassan, H., & Nassef, M. (2019). Automated language essay scoring systems: A literature review. *PeerJ Computer Science, 5,* e208. https://doi.org/10.7717/PEERJ-CS.208

Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgement and Decision Making*, *11*(4), 401–415. https://journal.sjdm.org/15/151110a/jdm151110a.pdf

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, augmented and mixed reality: Applications and case studies* (pp. 476–489). Lecture Notes in Computer Science, vol. 11575. Springer. https://doi.org/10.1007/978-3-030-21565-1_32

Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2010). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, *4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kamath, U., & Liu, J. (2021). *Explainable artificial intelligence: An introduction to interpretable machine learning*. Springer. https://doi.org/10.1007/978-3-030-83356-5

Khosravi, H., Buckingham Shum, S., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., 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

Kim, P. H., Ferrin, D. L., Cooper, C. D., & Dirks, K. T. (2004). Removing the shadow of suspicion: The effects of apology versus denial for repairing competence- versus integrity-based trust violations. *Journal of Applied Psychology*, *89*(1), 104–118. https://doi.org/10.1037/0021-9010.89.1.104

Kim, T., & Song, H. (2021). How should intelligent agents apologize to restore trust? Interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, *61*, 101595. https://doi.org/10.1016/j.tele.2021.101595

Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '16), 7–12 May 2016, San Jose, CA, USA (pp. 2390–2395). ACM Press. https://doi.org/10.1145/2858036.2858402

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, Á., & Wang, X. (2018). Designing academic writing analytics for civil law student self-assessment. *International Journal of Artificial Intelligence in Education*, *28*, 1–28. https://doi.org/10.1007/s40593-016-0121-0

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI '15), 29 March–1 April 2015, Atlanta, GA, USA (pp. 126–137). ACM Press. https://doi.org/10.1145/2678025.2701399

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *Proceedings of the 2013 IEEE Symposium on Visual Languages and Human-Centric Computing* (VL/HCC'13), 15–19 September 2013, San Jose, CA, USA (pp. 3–10). https://doi.org/10.1109/VLHCC.2013.6645235

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lim, B. Y., & Dey, A. K. (2010). Toolkit to support intelligibility in context-aware applications. *Proceedings of the 2010 ACM International Conference on Ubiquitous Computing* (UbiComp '10), 26–29 September 2010, Copenhagen, Denmark (pp. 13–22). ACM Press. https://doi.org/10.1145/1864349.1864353

Lins de Holanda Coelho, G., Hanel, P. H. P., & Wolf, L. J. (2020). The very efficient assessment of need for cognition: Developing a six-item version. *Assessment*, *27*(8), 1870–1885. https://doi.org/10.1177/1073191118793208

Long, Y., & Aleven, V. (2013). Supporting students' self-regulated learning with an open learner model in a linear equation tutor. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (AIED '13), 9–13 July 2013, Memphis, TN, USA (pp. 219–228). Springer. https://doi.org/10.1007/978-3-642-39112-5_23

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87. https://doi.org/10.1177/1555343411433844

Matzat, U., & Snijders, C. (2012). Rebuilding trust in online shops on consumer review sites: Sellers' responses to user-generated complaints. *Journal of Computer-Mediated Communication*, *18*(1), 62–79. https://doi.org/10.1111/J.1083-6101.2012.01594.X

McAuley, E., Duncan, T., Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for* Exercise *and Sport*, *60*(1), 48–58. https://doi.org/10.1080/02701367.1989.10607413

Meuwissen, M., & Bollen, L. (2021). *Transparency versus explainability in AI*. https://doi.org/10.13140/RG.2.2.27466.90561

Möhlmann, M., & Zalmanson, L. (2017). Hands on the wheel: Navigating algorithmic management and Uber driver's autonomy. *Proceedings of the 38th International Conference on Information Systems* (ICIS 2017), 10–13 December 2017, Seoul, South Korea. https://www.researchgate.net/publication/319965259

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*, *11*(3–4), 1–45. https://doi.org/10.1145/3387166

Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., & Klein, G. (2019). Explanation in human–AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *DARPA XAI Literature Review.* https://doi.org/10.48550/arXiv.1902.01876

Mueller, S. T., Veinott, E. S., Hoffman, R. R., Klein, G., Alam, L., Mamun, T., & Clancey, W. J. (2021). Principles of explanation in human–AI systems. *Proceedings of the 35th Conference on Artificial Intelligence* (AAAI-21), 8–9 February 2021, Virtual. https://doi.org/10.48550/arXiv.2102.04972

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, 53*(4), 914–931. https://doi.org/10.1111/bjet.13232

Nazaretsky, T., Cukurova, M., & Alexandron, G. (2022). An instrument for measuring teachers' trust in AI-based educational technology. *Proceedings of the 12th International Conference on Learning Analytics and Knowledge* (LAK '22), 21–25 March 2022, Online (pp. 56–66). ACM Press. https://doi.org/10.1145/3506860.3506866

Ooge, J., Kato, S., & Verbert, K. (2022). Explaining recommendations in e-learning: Effects on adolescents' trust. *Proceedings of the 27th International Conference on Intelligent User Interfaces* (IUI '22), 22–25 March 2022, Helsinki, Finland (pp. 93–105). https://doi.org/10.1145/3490099.3511140

Papenmeier, A., Englebienne, G., & Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust. *Proceedings of the 28th International Joint Conference on Artificial Intelligence* (IJCAI'19), 10–16 August 2019, Macao, China (pp. 94–100). https://doi.org/10.48550/arXiv.1907.12652

Prahl, A., & van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, *36*(6), 691–702. https://doi.org/10.1002/FOR.2464

Prinsloo, P., & Slade, S. (2018). Mapping responsible learning analytics: A critical proposal. In B. H. Khan, J. R. Corbeil, & M. E. Corbeil (Eds.), *Responsible analytics and data mining in education: Global perspectives on quality, support, and decision-making*. Routledge. http://oro.open.ac.uk/55827/

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

Rawal, A., McCoy, J., Rawat, D. B., Sadler, B. M., & St. Amant, R. (2022). Recent advances in trustworthy explainable artificial intelligence: Status, challenges and perspectives. *IEEE Transactions on Artificial Intelligence*, *3*(6), 852–866. https://doi.org/10.1109/TAI.2021.3133846

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD '16), 13–17 August 2016, San Francisco, CA, USA (pp. 1135–1144). ACM Press. https://doi.org/10.1145/2939672.2939778

Rosenthal, S. L., & Dey, A. K. (2010). Towards maximizing the accuracy of human-labeled sensor data. *Proceedings of the 15th International Conference on Intelligent User Interfaces* (IUI '10), 7–10 February 2010, Hong Kong, China (pp. 259–268). ACM Press. https://doi.org/10.1145/1719970.1720006

Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 5–22). Springer. https://doi.org/10.1007/978-3-030-28954-6_1

Sclater, N. (2016). Developing a code of practice for learning analytics. *Journal of Learning Analytics*, *3*(1), 16–42. https://doi.org/10.18608/jla.2016.31.3

Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, *6*(3), 11–19. https://doi.org/10.18608/jla.2019.63.3

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

Slade, S., & Prinsloo, P. (2013). Learning analytics: Ethical issues and dilemmas. *American Behavioral Scientist*, *57*(10), 1510–1529. https://doi.org/10.1177/0002764213479366

Snijders, C., Bober, M., & Matzat, U. (2017). Online reputation in eBay auctions: Damaging and rebuilding trustworthiness through feedback comments from buyers and sellers. In B. Jann & W. Przepiorka (Eds.), *Social dilemmas, institutions, and the evolution of cooperation* (pp. 421–444). De Gruyter Oldenbourg. https://doi.org/10.1515/9783110472974-020

Tzimas, D., & Demetriadis, S. (2021). Ethical issues in learning analytics: A review of the field. *Educational Technology Research and Development*, *69*, 1101–1133. https://doi.org/10.1007/S11423-021-09977-4

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

Warren, G., Keane, M. T., & Byrne, R. M. J. (2022). *Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in XAI*. https://doi.org/10.48550/arXiv.2204.10152

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94–109. https://doi.org/10.1016/J.COMPEDU.2016.05.004

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

Yang, L. W., Aggarwal, P., & McGill, A. L. (2020). The 3 C's of anthropomorphism: Connection, comprehension, and competition. *Consumer Psychology Review*, *3*(1), 3–19. https://doi.org/10.1002/arcp.1054

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19), 4–9 May 2019, Glasgow, Scotland, UK (pp. 1–12). ACM Press. https://doi.org/10.1145/3290605.3300509

Zumbrunn, S., Marrs, S., & Mewborn, C. (2016). Toward a better understanding of student perceptions of writing feedback: A mixed methods study. *Reading and Writing*, *29*, 349–370. https://doi.org/10.1007/S11145-015-9599-3