Aug 10th, 12:00 AM

Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring

Philipp Hartmann
University of Goettingen, philipp.hartmann@uni-goettingen.de

Sebastian Hobert
University of Goettingen, shobert@uni-goettingen.de

Matthias Schumann
University of Goettingen, mschuma1@uni-goettingen.de

Trust, but Verify! - An Empirical Investigation of Students' Initial Trust in AI-Based Essay Scoring

Completed Research

Philipp Hartmann
University of Goettingen
philipp.hartmann@uni-goettingen.de

Sebastian Hobert
University of Goettingen
shobert@uni-goettingen.de

Matthias Schumann
University of Goettingen
mschuma1@uni-goettingen.de

Abstract

AI is becoming increasingly important in supporting education. Nowadays, AI-based systems can score essays in high-stakes exams not only by comparing words but also by evaluating content. However, for AI-based essay scoring systems to be used, they must be trusted. Based on a scenario-based experiment with 260 students at a German university, we were able to show that their initial trust in AI-based essay scoring systems is significantly lower than in human examiners. Human control of AI-scoring can partially reduce the negative effect. The perceived system characteristics and the personality traits of the students are important factors which positively influence trustworthiness and trust, respectively. Furthermore, we could show that the more complex the essay scoring is perceived, the less trustworthy the AI-based system is classified. No influence could be seen regarding the relevance of the scoring for the students, their AI-experience and technology affinity.

Keywords

Trust, Education, Essay Scoring, Artificial Intelligence.

Introduction

Digital education has enjoyed growing popularity for years. This effect is additionally strengthened by increasing offers for Open Education Resources and MOOCs (Impey and Formanek 2021). The idea behind digital educational offerings is manifold. While participants can flexibly access learning resources regardless of location, educational institutions can expand their offerings. Capacity is no longer tied to premises, thus reducing the cost per participant (Yusuf and Al-Banawi 2013). However, this primarily concerns fixed costs, not variable costs. For example, with an increased number of participants there comes a potentially increased effort regarding individual support and the scoring of exams (Balfour 2013). While AI-based chatbots are already being used to handle individual support (Hobert 2019), knowledge assessment has mostly been handled with closed question types (Hewlett and Kahl-Andresen 2014). Since effective knowledge assessment is not possible with closed questions alone, open question types are increasingly used to examine higher taxonomy levels according to Bloom et al. (1956) (Birenbaum et al. 1992). Yet, since scoring open question is very costly, the use of AI in formative and summative assessments is more and more applied to solve this problem (Attali and Burstein 2006; Castellanos-Nieves et al. 2011; Richardson and Clesham 2021). Nowadays, AI-based essay scoring systems do not just allow simple word comparisons but also fully-automatic content and logic checking of whole essays (Pearson 2019). Furthermore, there are also benefits for examinees such as less scoring time and the removal of human bias (Richardson and Clesham 2021). Despite past efforts, AI-based scoring systems have so far failed to build trust among examiners and examinees (Kumar and Boulanger 2020; Richardson and Clesham 2021). Previous research on general AI-based services has shown that imperfect algorithms reduce trust and thus

acceptance (Kocielnik et al. 2019). Hence, when it comes to educational issues, students have more trust in people they know in the field than in the technologies being used (Richardson and Clesham 2021). This may be because AI-based essay scoring has its limitations such as the dependence on training data (Kumar and Boulanger 2020). Especially when examinees have to give their own opinion or a freely chosen example, AI reaches its limits. User trust is a particularly important but multifaceted construct here, influencing acceptance and thus usage (Wu et al. 2011). In the following, we will therefore investigate which factors influence an examinee's trust in AI-based scoring systems. In this context, trust in a relationship depends on three dimensions, namely the trustor (examinee), the trustee (AI-based essay scoring system), and the environment or situation (high-stakes exams), which are determined by different factors (Mayer et al. 1995; Siau and Wang 2018). While previous research has often focused on the trust of active users, we will look at the trust of passive users, who do not use the system themselves but are affected by its decisions. Thereby, trust is considered a dynamic system that consists of an individual basic trust (initial trust) as well as a trust that develops during the interaction (continuous trust) (Siau and Wang 2018). Since the use of AI-based essay scoring in high-stakes exams is still in its infancy, we will focus on initial trust. Initial trust describes the first contact between the two parties and is crucial for supporting the adoption of new technology. It is based on pre-implementation expectations (Li et al. 2008).

In the following, we will examine the factors that influence the examinee's initial trust in AI-based scoring systems using a scenario-based questionnaire study. Scenario 1 describes a semi-automatic system in which AI serves as a decision support system for a human scoring. Scenario 2 describes an automatic scoring system in which humans are no longer involved.

Related Research and Hypotheses Development

Most commonly, trust is defined as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or confront that other party" (Mayer et al. 1995). Although this definition deals with interpersonal trust, it can be transferred and adapted to the area of technology and AI-use. In the following, we will discuss the above-mentioned dimensions established by Mayer et al. (1995).

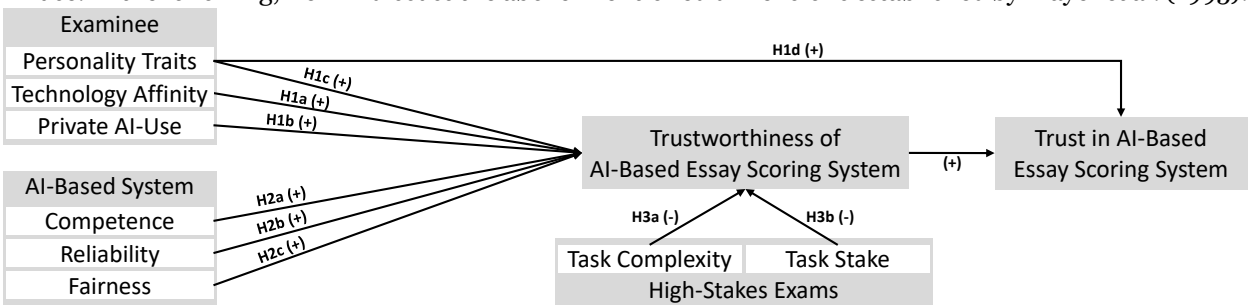

Figure 1. Trust Model Used for this Research Study

Examinee (The Trustor)

In our model, the examinees take on the role of human trustors. Each trustor has an individual propensity, i. e., willingness, to trust (Mayer et al. 1995). It is based on a generalization of various unique experiences (Lee and See 2004). The propensity to trust can be subdivided into ability- and personality-based factors (Siau and Wang 2018). Ability-based factors are grounded on information and knowledge about the trustee as well as on prior experiences and help to form predictions about the system's behavior. Since there is no comparable system in the context under investigation, the trustors do not have any information or knowledge from prior use of AI-based essay scoring systems. Therefore, this aspect is examined using the students' overall technology affinity and experience with other AI-based services (e.g., virtual assistants like Amazon Alexa or Apple's Siri). Former research showed that a high technology affinity promotes an increased tendency to actively approach and thus trust new technologies (Franke et al. 2019). We follow this argumentation and expect that a similar impact exists through the use of other AI-based services because experience with AI-based services in a private environment promotes understanding / reputation and hereby trust in other areas of use (Bao et al. 2021). Personality-based factors reflect the trustor's personality traits (Oleson et al. 2011). Prior research describes trust-related personality traits as the basis

for general trust before having information on a particular trustee (Mayer et al. 1995; Siau and Wang 2018). Especially in case of initial use, without sufficient information for a cognitive evaluation of the system, different personality traits (e.g., agreeableness) influence the emotional response to the system (Bao et al. 2021; Madsen and Gregor 2000). We assume that a higher agreeable personality trait leads to a higher trustworthiness of the AI-based scoring system as well as a higher overall trust in the AI-based scoring.

H1a: A higher technology affinity leads to a higher expected trustworthiness of the AI-based essay scoring system.
H1b: A higher experience in private use of AI leads to a higher expected trustworthiness of the AI-based essay scoring system.
H1c: A higher agreeable personality trait leads to a higher expected trustworthiness of the AI-based essay scoring system.
H1d: A higher agreeable personality trait leads to a higher trust in the AI-based essay scoring system.

AI-Based Essay Scoring System (The Trustee)

While the trustor's characteristics express the general willingness to trust, the trustee's characteristics describe the belief in its trustworthiness (Siau and Wang 2018). In previous research, attempts were made to transfer human attributes to AI. The factors ability (performance), benevolence (purpose), and integrity (process) are the basis for the trustee, as defined by Mayer et al. (1995) and adapted by Lee and See (2004). Since we are focusing on initial trust and no such system has been used with the participants so far, we will formulate the factors as expectations in the following. The expected performance describes the domain-specific skills and competences of the trustee (Mayer et al. 1995). It refers to the ability to achieve the trustor's goals in a specific task and situation and influences the expected trustworthiness (Lee and See 2004). The assumption is that highly competent trustees are more likely to perform delegated tasks satisfactorily on behalf of the trustor, without the need for control. In our context, examinees expect the exam to be scored by a person who is highly competent in the relevant domain (e.g., the lecturer). We assume that higher expected competence of the AI-based system leads to higher expected trustworthiness. The factor 'process' describes the perception that the trustee follows predefined joint principles that aim at promoting reliable action on the part of the trustee. Therefore we will focus on reliability. The experiences from previous actions are an important indication of the trustee's reliability. These experiences do not have to be made by the trustees themselves but can also arise from communication through others. Previous research has shown that merely the expected level of integrity is important and not why the perception exists (Mayer et al. 1995). Hereby, the factor does not describe a task-specific property, but a character property of the trustee (Lee and See 2004). In our case, the goal of the AI-based system is the proper scoring of essays in high-stakes exams. For examinees, it is therefore important that the AI performs the scoring reliably. So, we hypothesize that higher expected reliability of the AI-based system leads to higher expected trustworthiness. The factor 'purpose' shows the extent to which a trustee acts in the interests of the trustor and puts aside his own interests. Thereby a positive attitude by the trustee towards the trustor is assumed (Mayer et al. 1995). In the domain of IS, the factor focuses on the original intention for the development and also addresses the task that is to be accomplished (Lee and See 2004). Active users (examiners) and passive users (examinees) may have varying purposes. The examinee's goal is a fair assessment of the individual performance. An assessment can be considered fair if it correctly measures the individual's knowledge and also classifies it in relation to other examinees (Tierney et al. 2011). The system thus has the task of scoring essays without treating individual examinees unfairly. Therefore, we assume that higher expected fairness of the AI-based essay scoring leads to higher trustworthiness in the AI-based system.

H2a: A higher expected competence of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.
H2b: A higher expected reliability of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.
H2c: A higher expected fairness of the AI-based system leads to a higher expected trustworthiness of the AI-based essay scoring system.

High-Stakes Exams (The Environment)

The environment is determined by the task as well as cultural and institutional factors (Siau and Wang 2018). Institutional factors refer to the structural preconditions such as contracts, guarantees, or

regulations (Siau and Wang 2018). Cultural factors can be defined as the set of shared social norms associated with national or social differences (Lee and See 2004). Since we focus on students at a German university, we do not expect to observe any significant cultural as well as institutional differences in our sample. Consequently, these factors are not considered in the following. Despite constant human and AI-based factors, task-specific characteristics in the environmental context can influence trust levels (Mayer et al. 1995). Hence, the evaluation of the task characteristics plays an important role in the evaluation of trust. The risk of a task can be described by the task complexity (probability of failing) and the task stake (consequences of failing). Research showed that the type and severeness of the consequences have a significant effect on trustworthiness (Ashoori and Weisz 2019). Therefore a trustor will engage with a trustee if the level of trustworthiness surpasses the threshold of perceived risk (Mayer et al. 1995). In our context, we assess to what extent the scoring of high-stakes exams (e.g., final exams in mandatory courses) is relevant for the individual and can thus be considered a high-stakes task. Besides, we ask to what extent the scoring of essays is considered a complex task. We assume that the low degree of both variables leads to an increase in the trustworthiness of the AI-based essay scoring system.

H3a: A lower perceived task complexity of essay scoring leads to a higher expected trustworthiness of the AI-based essay scoring system.

H3b: A lower perceived task stake of essay scoring leads to a higher expected trustworthiness of the AI-based essay scoring system.

Research Design

Scenarios and Questionnaire Introduction

To analyze the hypotheses and thus answer the research question, students at a large German university were surveyed. The questionnaire was divided into two sections. Section 1 addressed the status quo of essay scoring in high-stakes exams and AI-independent items. Section 2 addressed the AI-use for essay scoring and AI-dependent items. To measure the hypotheses-related items, the participants were asked about their level of agreement with pre-formulated statements using a 6-point Likert scale (completely disagree (1) to completely agree (6)). Exceptions were the experience in private AI-use and the measurement of the personality traits. The experience in private AI-use was measured by frequency of use using a 6-point Likert scale (never (1) to daily (6)). The items of the personality trait were rated on a 5-point Likert scale (completely disagree (1) to completely agree (5)) and then compared with a benchmark for our target group. Overall, the questionnaire included 64 statements and questions.

In section 1, participants were asked about their demographic information, including age and gender. To ensure that all participants had a common knowledge concerning the scoring of essays at high-stakes exams, a short animated video about an exemplary exam situation and the associated scoring process was shown. Since the type and length of exam assignments can vary between courses, it was stated that only essays of approximately half to three-quarters of a page in length are included in the exam. The tasks included the reproduction, explanation, and transfer of the learned contents. The described scoring process represents the common procedure at German universities, which is carried out completely manually. Here, a four-eye principle was presented, which consists of a pre-scoring by a qualified employee and a final scoring by the professor in charge. In addition, the students were informed that this procedure entails longer scoring times, especially for larger courses. Based on this scenario, the AI-independent items were collected first. These included the trust in the described manual scoring process as well as an estimation of the expected scoring accuracy. For the trustor characteristics, the personality trait was queried using the German adaption of the Big Five Personality Traits Taxonomy (John et al. 2008), focusing on the trust-facet (dimension agreeableness). The ability factors were measured by using the students' technology affinity and individual experience in using AI-based services. The technology affinity was examined by employing the ATI-scale, consisting of a standardized questionnaire covering 9 items about engaging or avoiding technology interaction (Franke et al. 2019). The individual experience was assessed based on the frequency of use of voice assistants, facial recognition, and individual recommendation systems. Section 1 closed with the environmental factors, using the task complexity and task stakes.

At the beginning of section 2, the participants were randomly divided into two groups to investigate two scenarios in order to measure the influence of human scoring in our study. Both groups were shown an almost identical video. In the beginning, the participants were informed that the former described scoring

process can be shortened to a few days by using AI. The participants in *scenario 1* were told that the AI only takes over the pre-scoring and that the professor spot-checks this pre-scoring. The participants in *scenario 2* were told that the AI would take over the whole scoring automatically, without a spot-check by the professor. Following this, both groups were again identically given a brief description of how an AI works. The students were informed that a previously defined level of expectations is used for the scoring by the AI-based system. In addition, the system learns from previous exam scorings whose answers were assessed as partially or completely correct. The knowledge generated from the past scorings is then applied to the current scoring. Subsequently, it was explained that the comparison does not only take place on a word basis, but also considers synonyms, word combinations, and negations to guarantee a check of the content beyond sentences. Finally, it was pointed out that the AI can also make mistakes, but that human examiners also make mistakes to a comparable extent. Based on this video, the AI-dependent items were collected. First, the AI-based system factors as well as the items about the trustworthiness of the AI-based essay scoring system were surveyed. Second, similar to section 1, students were again asked about their trust in the described scoring process as well as their expected scoring accuracy. In addition, students were asked whether they would attend an exam review more often if the AI-based system was used instead of human scoring. In a final step, students had the opportunity to provide further comments on the AI-based exam scoring in a short text field.

Data Collection and Pre-Processing

Factor	**Items**
Competence (CA = 0.765)	The AI-based system has in-depth knowledge of scoring exams.
	The scoring results of the AI-based system are as good as those of a highly competent person.
	The AI-based system correctly scores the exam answers I submit.
	The AI-based system uses all the knowledge and information at its disposal to score an exam.
Reliability (CA = 0.704)	The AI-based system works reliably.
	The AI-based system scores comparable exam answers of different exam participants equally.
	I can rely on the AI-based system to work flawlessly.
	The AI-based system scores the exam answers without contradictions.
Fairness (CA = 0.625)	I believe that an AI-based system would be used in my best interest.
	The AI-based system looks after my interests, not just those of the professor.
	During AI-based scoring, preference is given to individual examinees.
	The AI-based system ensures a fair scoring of the individual performance of examinees.
Task Complexity (CA = 0.563)	The scoring of exams is demanding.
	For the scoring of an exam task, one needs a specialized knowledge that exceeds the knowledge for the answering of the task.
	The optimal answer to an exam task is always unique.
	Errors rarely occur in the scoring of exams.
Task Stake (CA = 0.615)	The correct scoring of an exam is very important.
	The grade in an exam has a long-term impact on the student's life.
	I care about good grades.
	If I get a lower grade than expected, I don't think about it for very long.
Trustworthiness of AI-Based Essay Scoring System (CA = 0.800)	The AI-based scoring process is trustworthy.
	I would change one or more aspects of the scoring process to make AI-use trustworthy.
	The AI-based scoring process will result in a fair outcome for the examinees.
	Examinees need more information about how the AI-based system scores in order to trust the scoring process.
Trust in AI-Based Essay Scoring System (CA = 0.632)	I trust the AI-based scoring process of exams.
	I would like to keep an eye on the AI-based system during scoring.
	The exam reviews of the final scoring by examinees are needed to control the AI-based scoring.
	The AI-based system should be more controlled.
	For the AI-based system, its own interests (e.g., the lowest possible scoring effort) are paramount in the scoring process.

Table 1. Reliability Coefficients of the Factors and Items Used

The questionnaire was forwarded to students at a German university via multiple channels (e.g., e-mail, forum, personal addresses in classroom lectures). Participation was anonymous and voluntary. Vouchers

were raffled among all participants who completed the questionnaire. A weighting of the participants according to gender, age, or other criteria was not carried out. A total of 330 students took part in the survey. Due to the use of incentives, it can be assumed that some participants did not show the required seriousness. To reduce disruptive effects, we tried to remove these participants by identifying outliers in the processing time. This leaves a data sample of 260 participants, of whom 51.92 % were male and 48.08 % female. Their age varied between 18 and 35 years (MD = 21.85; SD = 2.69). 51.54 % of the participants were shown scenario 1. Scenario 2, on the other hand, was seen by 48.46 %.

In the selection of items used, we drew on existing and scientifically tested items, which were adapted to the subject of manual and (semi)-automatic AI-based essay scoring. To assess the fit of the model with our collected data, a confirmatory factor analysis was conducted for the existing scales, while an exploratory factor analysis was used for the others. Since we used the already validated ATI-score (Franke et al. 2019) to measure the technology affinity and the Big Five Personality Traits Taxonomy (John et al. 2008) to measure the personality trait, we excluded these items from the factor analysis. The sample has a KMO-value of 0.840 and can be considered suitable for factor analysis (Hair et al. 2018; Kaiser 1974). Due to cross-loadings and poor factor loadings, we removed certain items to ensure construct reliability (highlighted in gray in Table 1). For the remaining items, we conducted tests for convergent validity by determing the composite reliability (CR) and the average variance extracted (AVE). The values for both indicators are above the critical values and therefore at an acceptable level (Hair et al. 2018). The items used and the associated Cronbach's Alpha (CA) values for the identified factors are listed in Table 1.

Results

Descriptive Analysis

Regarding the scenario-independent factors, the following values were obtained. For trust in manual scoring, the participants stated that they trusted the scoring in principle (MD = 3.86; SD = 0.83). Among the personal trait factors, above-average values can be observed for the trust-facet (MD = 3.76; SD = 0.67). When dealing with technologies, the ATI shows a mean average technology affinity (MD = 3.65; SD = 1.08). Greater differences are evident in the use of AI-based services. For example, when using voice assistants, 60.8 % said that they use them only once a month or fewer, whereas only 20.4 % use them (almost) daily. Regarding the use of facial recognition, 35.0 % indicated infrequent use, while 62.3 % use it (almost) daily. For the use of individual recommendations, the proportions are 32.3 % and 28.1 %. High values can be observed for the environmental factors. The participants rated the scoring of exams as complex (MD = 4.94; SD = 0.84) and important (MD = 4.89; SD = 0.86). T-tests show no significant difference between the participants of the scenarios. The results for the scenario-dependent factors are shown in table 2.

Factor	**MD (S1/S2) SD (S1/S2)**	**T-Value (df = 258)**	**Factor**	**MD (S1/S2) SD (S1/S2)**	**T-Value (df = 258)**
Competence	MD (4.19 /4.01) SD (0.82 / 0.82)	T = 1.698 *	Trustworthiness of AI-Based Essay Scoring System	MD (4.17 / 3.94) SD (0.94 / 1.07)	T = 1.891 *
Reliability	MD (3.78 / 3.55) SD (0.84 /0.85)	T = 2.167 *	Trust in AI-Based Essay Scoring System	MD (3.63 / 3.26) SD (1.09 / 0.97)	T = 2.863 **
Fairness	MD (4.37 / 4.21) SD (0.77 / 0.86)	T = 1.572	*** p<0.001; ** p<0.01; *p<0.05		

Table 2. Descriptive Results of AI-Related Factors

Furthermore, in the semi-automatic scenario, 59.7 % of the respondents indicated that they would be more likely to attend an exam review if AI was used. In the automatic scenario, the proportion was 71.4 %. The statistical analysis showed that there is a significantly higher percentage of students in scenario 2 who would participate in the review (p<0.001). The expected accuracy of the manual scoring was reported by the participants with a mean value of 83.71 % (SD = 9.63 %). Surprisingly, no significant difference to the AI-based scoring can be observed. Thus, the participants indicated comparable accuracies for the semi-

automatic scoring from scenario 1 (MD = 83.99 %; SD = 12.21 %) and for the automatic scoring from scenario 2 (MD = 82.61 %; SD = 13.00 %). T-tests show no significant difference between the scenarios and in comparison to the human scoring.

Statistical Analysis

To test the postulated hypotheses, the statistical software Stata SE was used. Before we conducted the structural equation model, the assumptions were checked. The multivariate normality was checked using the Mahalanobis distance. No further outliers were observed. For multicollinearity, all VIF-values and tolerances were at an acceptable level (Hair et al. 2018). The structural equation model was estimated using the maximum likelihood estimation and model fit indices were determined. The coefficients and the corresponding significance levels can be seen in Figure 2. Overall, the model has an acceptable to good fit for different quality indices. So the values for RMSEA (0.080), CFI (0.971), TLI (0.930), and SRMR (0.023) are all at an acceptable level (Hair et al. 2018).

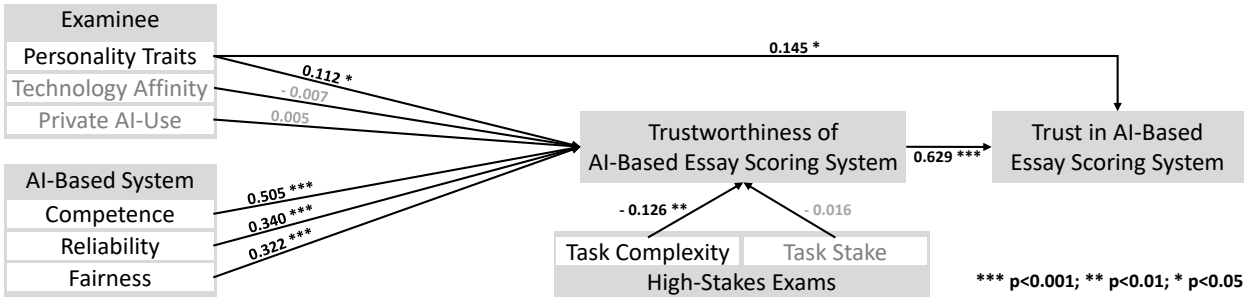

Figure 2. Results of the Structural Equation Model

Discussion and Implication

Examinee (The Trustor)

Concerning the trustor, we investigated ability and personality traits. Hypotheses 1a and 1b dealt with the influence of ability on the trustworthiness of the AI-based scoring system. These hypotheses could not be confirmed. Hypotheses 1c and 1d dealt with the influence of personality traits on the trustworthiness of AI-based scoring systems and towards trust in AI-based scoring. These hypotheses could be confirmed.

We were thus able to show that previous experience with AI-based services and technology affinity do not influence the trustor's trust propensity towards the AI-based essay scoring system. One possible reason for this could be that the examinees were not active but merely passive users since they did not directly interact but were only confronted with the outputs of the system. As for the personality traits, we observed significant influences concerning the trust-facet (dimension agreeableness) and were able to confirm the existing research results. Since the participants can be described as young and educated, the level of the trust-facet was considered in relation to this benchmark (Danner et al. 2019). We were able to show that participants with an above-average level of the trust-facet showed a higher perceived trustworthiness of the AI-based system and a higher trust in AI-based scoring. Participants whose level of the trust-facet is lower than the benchmark showed a lower level of trustworthiness towards the AI-based system compared to the benchmark. Since personality traits are formed over a long period of time based on individual experience, it is not possible to exert any short-term influence to increase trust propensity towards AI-based systems.

AI-Based Essay Scoring System (The Trustee)

The trustee characteristics influence how the trustee is perceived by the trustor and the amount of trustworthiness assigned to him. We tested the factors of competence, reliability, and fairness. The hypotheses H2a to H2c were all confirmed. Fairness was rated as equally high in both scenarios. Thus, the additional spot-checks by the professor did not lead to any changes. For reliability, a significant difference was observed between the scenarios: the value of the semi-automatic is higher than that of the automatic scoring. The system seems to have a lower overall reliability, which can be partly compensated by the control

of the professor. We observed a significant difference regarding competence. Here, too, the semi-automatic process is perceived as significantly more competent. So, the role of AI in the scoring process has an important influence. Indeed, previous research has shown higher trust in human deciders than in automatic AI-based systems, especially for important decisions (Ashoori and Weisz 2019). In our initial scenario, we described the task as the reproduction, explanation, and transfer of learned content. The system does not seem to be trusted to possess a competence equal to that of humans. One reason for this may be the task of explanation and transfer, whose answers cannot be classified into right or wrong in the level of expectations and are thus difficult to teach to the system. The closer the answer to a firmly defined level of expectations, the higher the quality of the scoring. Here, a lack of transfer to individual examples could be a possible cause for the lower perceived competence. Previous research has also shown that students still trust the people they associate with the activity more than systems (Elson et al. 2021; Richardson and Clesham 2021). This may show a negative image of AI since these results are in contradiction to the expected accuracy, where we could not identify any differences between the manual scoring and the scenarios. Overall, system-related factors represent the most important influence on trustworthiness and thus trust over AI-based essay scoring. As a result, an attempt could be made to increase fairness and reliability through the transparent implementation of protocols for proper essay scoring.

High-Stakes Exams (The Environment)

For the environment characteristics, we focused on the task-related factors. Hypothesis 3a, in which we stated a negative relationship between the perceived task complexity and the trustworthiness of the AI-based essay scoring system, was confirmed. Thus, the task was perceived as very complex, with a significant negative influence on trustworthiness confirming the results of previous research (Ashoori and Weisz 2019). Hypothesis 3b, assuming that a high perceived relevance of the scoring also influences trustworthiness, could not be confirmed. Although the correct scoring of exam tasks was also assigned as important, this did not have any significant influence on trustworthiness in our case. One reason for this could be a good task-AI fit. For the trustor, the appropriate completion of the task is of primary importance. If a trustee, in our case the AI-based scoring system, is in sum rated as competent to perform the assigned task, it may not matter how relevant the task is to the examinee. An indication of a good task-AI fit may be that in both scenarios the scoring was perceived as fair and the system as competent. Depending on the task complexity, we recommend to design the use of the AI-based system appropriately. We therefore suggest, that for complex tasks, the semi-automatic use of AI as a decision support system should be considered. Thus, human control can increase perceived competence and ensure a better task-AI fit.

Limitation

As with any similar quantitative questionnaire study, we are aware of various limitations. First, our model attempts to explain trust in AI-based (semi-)automatic essay scoring in high-stakes exams through the trustworthiness of the AI-based system and personal characteristics. By conducting a factor analysis, we combined multiple existing and newly created items into the postulated factors. Since the subject of trust in AI-based essay scoring is quite new, we cannot assure that our results are complete. Thus, many assumptions of the model under consideration are based on the trustor as an active user. In our case, however, the examinees represent passive users who are just confronted with the results. In this area, prior research is still in its infancy. Second, we primarily tried to use existing, valid items, which were translated and adapted to our context and target group. As a result, important linguistic facets may have been lost. Additionally, a narrow set of factors (competence, reliability and fairness) was selected for the AI-based system characteristics, so that possible dimensions may not have been considered. Furthermore, the personal traits were measured using the Big Five Personality Traits Taxonomy (John et al. 2008). The determination of a complete personality profile can comprise up to 240 items and is therefore difficult to implement in the context mentioned (Costa, Jr. and McCrae 2000). Here, the focus was placed only on the trust-facet as part of the dimension agreeableness, which is measured by 4 items. Overall, other personality traits could also influence trust. Future research should therefore focus on additional personality traits to provide further insights into the influence on the trustworthiness of and trust in AI-based services. Third, new items were developed for individual constructs. In this respect, the factor analysis revealed possibilities for improvement. The difficulty in operationalizing trust is that different items are reliable for measuring trust in AI and trust in humans. It is therefore difficult to formulate a uniform set of items that allows direct comparisons of humans and AI. Furthermore, there is room for improvement in the scales for the

environmental factors. To be able to validly assess trustworthiness and thus trust in different situations, it must first be possible to clearly define the situational context. Here, our Cronbach's Alpha-values for task complexity and task stake still show potential for improvement. Fourth, as mentioned before, the survey was only conducted at one university in Germany. Although the culture and scoring process among German universities is quite similar, there may occur regional as well as national differences. The results of this study can therefore only represent a starting point for further research and still needs to be verified regarding its generalizability.

Conclusion

For a long time, the use of AI-based services was only possible to a limited extent due to technical limitations. The benefits of AI-based services depend on the available database with which the system is trained. Due to the growing availability of large data sets, this limit is gradually being overcome, so that AI-based services are increasingly being used in different areas. This applies to the education sector, where students in a more and more digitized education are enabled to receive individual support even in large digital courses. However, previous research has shown that the use of AI-based systems depends on the users' trust in them. We could show that especially in situations perceived as complex, such as high-stakes exams, the trustworthiness of the AI-based system is not high. Thus, the trust in automatic AI-based essay scoring is still significantly below the trust in manual scoring. This lack of trust can be partially reduced by using AI as a decision support system for human decision makers. In the case of the trustor characteristics, the individual trust-facet of the personality traits is an important factor for the trustworthiness of and the trust in the AI-based system. The trustworthiness of these systems heightens with increasing expectations concerning competence, reliability and fairness. No influence was found regarding the technical abilities or the relevance of the task. It is also interesting to note that despite the differences in trust, no significant differences in the expected scoring accuracy were observed between the manual and the two scoring processes in the scenarios. Here, there seems to be an unfounded skepticism towards the use of AI, which may be due to a general caution in society. As AI-based scoring systems become more widespread, the need for future research arises as well. Hence, aspects such as continuous trust can be investigated through regular use and a connection between intention and behavior can be examined (Ajzen 1991).

References

Ajzen, I. 1991. "The theory of planned behavior," *Organizational behavior and human decision processes* (50:2), pp. 179-211.

Ashoori, M., and Weisz, J. D. 2019. "In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes,"

Attali, Y., and Burstein, J. 2006. "Automated Essay Scoring With e-rater® V.2," *Journal of Technology, Learning, and Assessment* (4:3).

Balfour, S. P. 2013. "Assessing Writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review™," *Research & Practice in Assessment* (8), pp. 40-48.

Bao, Y., Cheng, X., de Vreede, T., and de Vreede, G.-J. 2021. "Investigating the relationship between AI and trust in human-AI collaboration," *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 607-616.

Birenbaum, M., Tatsuoka, K. K., and Gutvirtz, Y. 1992. "Effects of response format on diagnostic assessment of scholastic achievement," *Applied psychological measurement* (16:4), pp. 353-363.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., and Krathwohl, D. R. 1956. *Taxonomy of educational objectives: the classification of educational goals*: *Handbook I: cognitive domain*, New York, US: David McKay Co Inc.

Castellanos-Nieves, D., Fernández-Breis, J. T., Valencia-García, R., Martínez-Béjar, R., and Iniesta-Moreno, M. 2011. "Semantic Web Technologies for supporting learning assessment," *Information Sciences* (181:9), pp. 1517-1537.

Costa, P. T., Jr., and McCrae, R. R. 2000. "Neo Personality Inventory," *Encyclopedia of psychology* (5), pp. 407-409.

Danner, D., Rammstedt, B., Bluemke, M., Lechner, C., Berres, S., Knopf, T., Soto, C. J., and John, O. P. 2019. "Das big five Inventar 2," *Diagnostica*.

Elson, J. S., Derrick, D. C., and Merino, L. A. 2021. "An Empirical Study Exploring Difference in Trust of Perceived Human and Intelligent System Partners," *Proceedings of the 54th Hawaii International Conference on System Sciences*, pp. 136-145.
Franke, T., Attig, C., and Wessel, D. 2019. "A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale," *International Journal of Human–Computer Interaction* (35:6), pp. 456-467.
Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. 2018. *Multivariate data analysis*, Englewood Cliffs, NJ: Prentice Hall.
Hewlett, C., and Kahl-Andresen, A. 2014. "Prüfungsökonomie statt Prüfungsqualität?" *Berufsbildung in Wissenschaft und Praxis* (14:3), pp. 6-9.
Hobert, S. 2019. "Say hello to 'coding tutor'! design and evaluation of a chatbot-based learning system supporting students to learn to program," *Proceedings of the 40th International Conference on Information Systems*.
Impey, C., and Formanek, M. 2021. "MOOCS and 100 Days of COVID: Enrollment surges in massive open online astronomy classes during the coronavirus pandemic," *Social sciences & humanities open* (4:1).
John, O. P., Naumann, L. P., and Soto, C. J. 2008. "Paradigm shift to the integrative Big Five trait taxonomy: History, measurement, and conceptual issues," *Handbook of personality: Theory and research*, pp. 114-158.
Kaiser, H. F. 1974. "An index of factorial simplicity," *Psychometrika* (39:1), pp. 31-36.
Kocielnik, R., Amershi, S., and Bennett, P. N. 2019. "Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems," *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-14.
Kumar, V., and Boulanger, D. 2020. "Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value," *Frontiers in Education* (5).
Lee, J. D., and See, K. A. 2004. "Trust in automation: designing for appropriate reliance," *Human factors* (46:1), pp. 50-80.
Li, X., Hess, T. J., and Valacich, J. S. 2008. "Why do we trust new technology? A study of initial trust formation with organizational information systems," *The Journal of Strategic Information Systems* (17:1), pp. 39-71.
Madsen, M., and Gregor, S. 2000. "Measuring Human-Computer Trust," *Proceedings of the 11 th Australasian Conference on Information Systems* (53).
Mayer, R. C., Davis, J. H., and Schoorman, F. D. 1995. "An integrative model of organizational trust," *Academy of management review* (20:3), pp. 709-734.
Oleson, K. E., Billings, D. R., Kocsis, V., Chen, J. Y. C., and Hancock, P. A. 2011. "Antecedents of trust in human-robot collaborations," pp. 175-178.
Pearson. 2019. "PTE Academic Automated Scoring White Paper: Pearson Test of English Academix: Automated Scoring," Pearson Education Ltd (ed.), Pearson Education Ltd.
Richardson, M., and Clesham, R. 2021. "Rise of the machines? The evolving role of AI technologies in high-stakes assessment," *London Review of Education* (19:1), pp. 1-13.
Siau, K., and Wang, W. 2018. "Building trust in artificial intelligence, machine learning, and robotics," *Cutter business technology journal* (31:2), pp. 47-53.
Tierney, R. D., Simon, M., and Charland Julie. 2011. "Being fair: Teachers' interpretations of principles for standards-based grading," *The Educational Forum* (75:3), pp. 210-227.
Wu, K., Zhao, Y., Zhu, Q., Tan, X., and Zheng, H. 2011. "A meta-analysis of the impact of trust on technology acceptance model: Investigation of moderating influence of subject and context type," *International Journal of Information Management* (31:6), pp. 572-581.
Yusuf, N., and Al-Banawi, N. 2013. "The impact of changing technology: The case of e-learning," *Contemporary Issues in Education Research (CIER)* (6:2), pp. 173-180.