Influence of human-in-the-loop on the acceptance of AI-driven evaluation of essay questions by students.

Author: Kristians Balickis
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

ABSTRACT,

Artificial intelligence (AI) is developing rapidly in the various fields of our lives. The education field could benefit from AI-based systems as well. Introducing AI-based evaluation systems for essay questions in higher education could benefit universities and their staff. Yet this area is underexplored. The existing research focuses on the technical aspects of AI-based systems for grading and not on the human side. This research finds the acceptance levels of AI-based evaluation from a student perspective. This would be important since the acceptance of the evaluation process in universities highly correlates with their reputation.

A survey with Likert-type items was conducted. Participants answered the same questions three times for three different implementation scenarios of AI-driven evaluation. The first scenario is fully AI grading, and the two others include human-in-the-loop AI (HitAI). The result shows higher overall acceptance for scenarios with HitAI compared to a scenario with solely AI evaluation. Lastly, the discussion states limitations, and suggestions for future research. Practical implications for universities that planning to adopt such AI systems would be to keep humans in the loop.

Graduation Committee members:
Daniel Braun
Patricia Rogetzer

Keywords
AI Acceptance, Higher Education, Artificial Intelligence, Evaluation, Students, Essay Questions, Human in the Loop.

1. INTRODUCTION

Artificial Intelligence (AI) is developing rapidly ever since the influential paper by Turing in 1950. In his paper, Turing describes how a computer program can behave intelligently. In present-day, AI is a broad topic that finds applications in various segments of our lives (Haton, 2006). This project focuses on the education sector, specifically on the AI-driven grading of open/essay questions. An automated system for grading multiple-choice questions is not new. Evaluation of the essay questions, however, is a much more complex task and one that can be tackled with different approaches. Natural language processing (NLP) is a subfield of AI that deals with the processing of natural (human) language by machines. Other approaches in the context of evaluation of essay questions are Machine Learning (ML) algorithms that can assess answers based on a set of provided keywords. This paper however does not focus on the actual methods of evaluation but on perceptions of students that are potentially being graded using such a system.

2. RESEARCH OBJECTIVES AND RESEARCH QUESTIONS

The objectives of this research project are to measure the acceptance of AI-driven evaluation by students in higher education. It will be measured using theories from previous research done in this field, the theories will be adapted for the objectives of this research. The first objective is to measure the acceptance of AI-driven assessment without any human interaction (solely AI assessment). Then measure acceptance while considering HitAI approaches. Two different HitAI approaches will be considered. Following that it will be determined what approach will give the highest acceptance level. Appropriate data analysis methods will be used to determine that. One of the hypotheses of the research is that acceptance of the AI-driven evaluation will be higher in university students when the HitAI approach is used, compared to exclusively AI assessment.

Therefore, the research questions were formulated as follows:

RQ1: What is the acceptance level of solely AI-driven assessment of essay questions from a student perspective in higher education?

RQ2(a): What is the acceptance level of AI-driven assessment of essay questions in the view of students in higher education if teachers are given the last decision in the grading process?

RQ2(b): What is the acceptance if a teacher only reviews failed assignments and students are given the right to ask for a review from the teacher?

RQ3: Which human-in-the-loop approach will yield the highest acceptance level of AI-driven evaluation of essay questions in the view of students in higher education?

To answer the questions a survey will be conducted among current higher education students. The survey aims to understand general levels of acceptance when presented with the broad context of AI evaluation. Then check if the human intervention influences that level.

3. PRACTICAL AND ACADEMIC RELEVANCE

Extensive research was done through the years on the different kinds of evaluations. It is known for instance that teacher evaluations of the students are not free of bias, student evaluations are influenced by gender, race, personal traits, or likability (Lilly et al., 2022). In theory, an AI-driven approach could help to solve such problems. However, there is evidence that suggests that students would not necessarily prefer biased or bias-free methods. A study on the fairness of AI in higher education admission found that "both distributive and procedural AI fairness perceptions contribute to students' intention to raise the voice against the AI-driven admission system by protesting". This leads to tensions between the university and students, which may have negative effects on student performance and university reputation (Marcinkowski et al., 2020).

That raises the importance of considering the voice of the students and acceptance levels when applying the AI-driven systems, in the case of this research the AI-driven evaluations/grading.

Studies have shown that there is a correlation between fairness and acceptance of the grading process with a perceived reputation of the university, as was pinpointed in a review by Lilly et al. (2022) Therefore, this research can attempt to help universities with the adoption of AI-driven evaluation methods.

The universities might adopt AI-driven technologies for many reasons, the main reasons being reducing the time and resources needed for achieving certain tasks as found in the reviewed literature. In the context of university admissions, the institutions spend a considerable amount of human and financial resources on such a process (Marcinkowski et al., 2020). In medical education, professors are faced with time-consuming tasks of constructive feedback on assignments. However, it was found that an automated evaluation system would allow to grade student papers twice as fast while keeping the same number of teachers (Gierl et al., 2014).

More generally speaking the computer-based assessment provides cost and time reduction along with faster results (Terzis & Economides, 2011).

Moving forward there is an array of other possible advantages and a large number of applications as was summarized in the recent review (Zawacki-Richter et al., 2019).

Besides the perceived acceptance, this paper aims to find out if the human intervention (i.e., teacher intervention) will positively affect the acceptance levels. Zanzotto, (2019) in his article provides a viewpoint on human-in-the-loop AI (HitAI). One of the methods to keep humans in the loop is to give the last word to humans. In the context of this paper that means to give the last grading decision to the teacher while using the AI-driven system as an advisor. This can mean that teachers will grade every answer of the student as it is now while taking the AI into account or reviewing only answers that are graded as unsatisfactory by the AI. There could be other approaches to keeping humans in the loop as well. This approach might improve the acceptance levels in the view of students that are being evaluated. One of the aims of this paper then is to test that.

Lastly, talking about automated evaluation systems in general, there is a lack of research and significant challenges exist in adopting these systems (Ramesh & Sanampudi, 2022). This paper can contribute to a better understanding of the topic in general as well as improve the implementation of such a system.

Notably, no research was done on the perceived acceptance of AI-driven evaluation and the influence of the human intervention on that acceptance. That would be clearer in the next section about related literature. The research gap will be shown, considering that the study on the acceptance of the AI-driven evaluation in higher education is planned to be done soon by Sanchez-Prieto. However, the plan does not include the HitAI concept as it will become evident in the next sections.

4. LITERATURE REVIEW

From the reviewed literature it is evident that a small number of researchers focus on the object of this study. The study of the adoption and acceptance of AI in the education field constitutes an underexplored area of research (Sánchez-Prieto et al., 2020). Additionally, the automated (including the AI) systems for assessing essays or short answers are not developing rapidly. As was mentioned in the previous section, a lack of research and significant challenges exist in the adoption of such systems.

More to it, there is a lack of grading systems that are capable of grading open/essay questions and short answers. A recent review on automated scoring systems by Ramesh & Sanampudi, (2022) point out that research focused more on the non-content systems. That means that the grading system assesses the factors such as grammar or vocabulary while not considering the actual content or relevance of the answer. The amount of research papers that are focusing on the content is two times lower than non-content papers. Additionally, many ML models cannot differentiate the meaning of the word in various contexts. One word can have different meanings in different fields of study. Finally, no ML models exist that evaluate the relevancy of the answer to the question.

This might raise the question if students and universities even need AI or other automated approaches to grading and wheatear it is worth spending time and effort to develop such systems. The academic and practical relevance section describes the benefits for the university and its teachers that can be associated with using AI-driven grading. The students might experience benefits as well, for instance, the grading process might take less time compared to manual grading. Nevertheless, the perspective of students needs to be considered as the acceptance of the grading process correlates with the university's reputation (Lilly et al., 2022).

The study by Marcinkowski et al., (2020) addresses the perceptions of fairness by students when faced with an AI-driven system for university admissions. It was found that an AI-driven system was perceived to be much fairer than a human committee. However, the study was done with participants from only one university, thus other similar studies may contradict this.

Again, their findings are in line with the literature claiming that AI is perceived as a fairer agent (when compared to humans) in high-impact decisions (Araujo et al., 2020). In this case, the high-impact decision is university admission.

Assuming that grading exams/other assignments with open questions are high-impact decisions, students might have high acceptance levels of AI-driven evaluation.

There is little to no research about the influence of human in the loop (HitAI) on the acceptance of AI. As stated, before Zanzotto, (2019) provided a viewpoint on how to keep humans in the loop. That article is used as inspiration to design this study.

The literature on HitAI is not extensive and there are opposing opinions about this concept. Zanzotto, (2019) is enthusiastic about including humans and believes that this would be the future, not only by technical norms but also author believes it is morally correct to include the humans in the AI process. While (Cranor, 2008) writes about HitAI in a vastly different view, the author states that humans should be always kept out of the loop where possible. A good design of the AI should not require human intervention. This article however is not recent as opposed to Zanzotto's but it is also a highly popular one with citation counts close to 400. Additionally, Cranor describes strategies for building secure systems that humans can use. The system that automatically grades the exams or essays' primary function would not include being secure, yet it is worth pointing out one specific strategy. A system should be done in a way that is easily understood by humans to perform successfully. If it is not done correctly teacher might for instance not use the system at all and rely solely on his decisions. Which is not the intended outcome of implementing an AI-based grading system at the university.

Overall, it is evident that the literature is not extensive when it comes to the AI systems used in education. There is also a theoretical gap in this field as will become evident in the following section. Sanchez-Prieto's recent work points out these gaps as well, his future work will address this by conducting a study on the acceptance of AI in education from a student perspective.

This study however will fill the gap by addressing the acceptance of the AI system used for grading open questions, while considering the influence of HitAI.

5. THEORETICAL FRAMEWORK

This section includes the literature on theories used to measure the acceptance of technology. Following that the conceptual framework is constructed based on the previous theories.

5.1 Theoretical background

Upon reviewing the related literature on the acceptance of technology it was found that extensive research was made using the technology acceptance model (TAM). There exist multiple variations and extensions along with newer versions called TAM2 and TAM3(Cruz-Benito et al., 2019). Despite its popularity and use in many influential papers, this model might not be the best fit for this research. This model cannot define the acceptance of AI-driven evaluation, at least not in the original form.

The TAM model was adopted and changed to fit the research; this is the common practice when research is focusing on a specific technology. Moreover, one of the aims of this paper is to measure the acceptance with the HitAI concept while other models do not account for it. Although there is not much research on the student's acceptance of AI assessment yet, the attempts to develop a suitable model are there. Recently the expanded TAM was proposed, using the basis from the original model they expand it by adding new constructs to better relate to the topic. While they have a number of new and related constructs, they do not include the HitAI.

This research takes the model developed by Sánchez-Prieto et al., (2020) and modifies it to fit the context of the paper and reduce the scope while considering the HitAI.

Figure 1 shows the original TAM model proposed by Davis, (1989).

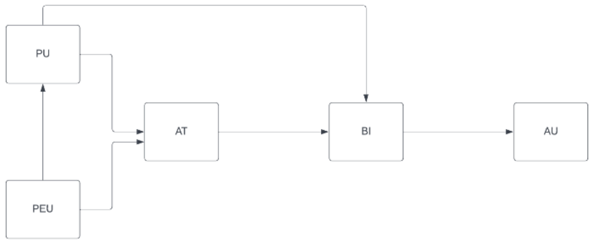


Figure 1. TAM (Source: Davis, 1989)

TAM describes the adoption of specific technology based on five constructs (Sánchez-Prieto et al., 2020):

- **Perceived Usefulness (PU):** A person's perception of the effect that the technology use has on his work performance

- **Perceived ease of use (PEU):** A person's assessment of the degree of effort needed to use the technology

- **Attitudes towards use (AT):** the feelings, opinions, favourable and unfavourable assessments about the use of technology

- **Behavioural intention (BI):** the level of willingness of the person to use the technology

- **Actual use (AU):** the frequency of use of the technology by a person.

PU and PEU are the main drivers of the model, they are conditions for attitudes towards user use, which is the condition of the BI which would be the factor leading to the actual use (Sánchez-Prieto et al., 2020).

The expansion was made to fit the object of the study. The proposition was made to include three more constructs to measure environmental pressure, (dis-)trust concerning the AIs, and natural opposition to the individual change. Explanation of constructs:

- **Social norm (SN):** social pressure to use AI assessment.

- **Trust (TR):** the willingness of individuals to rely on the AI assessment.

- **Resistance to change (RC):** opposition of individuals to move from status quo to AI assessment.

Figure 2 demonstrates the relations between different constructs including the three new ones **in bold**. RC construct is an inhibiting factor that negatively affects students' acceptance and is displayed in red colour in the model.

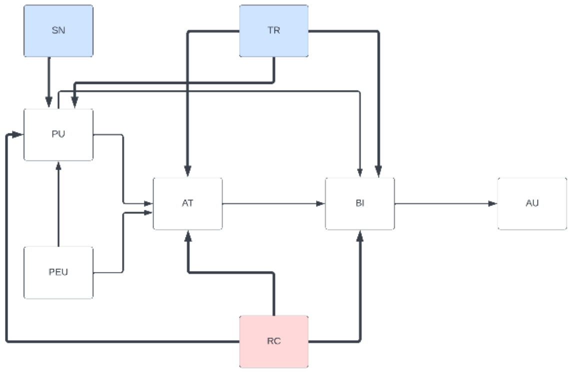


Figure 2. AI Assessment TAM (Source: Sánchez-Prieto et al., 2020)

Additionally, their work proposes the 30 Likert-type items to measure the constructs. The authors also stress that the topic might not be understood by all students and a small introduction would be needed to make sure everyone who participates in the survey would be on the same page. Lastly, a difference may exist between gender, age, a branch of knowledge, and experience with AI.

5.2 Development of the conceptual framework

This paper aims to find the acceptance level of the AI-driven assessment while considering HitAI's influence. This is done by composing a survey and measuring every construct for several different scenarios, where teacher presence in the assessment process would differ. Due to scope limitations and the nature of the research questions, the model is modified to reduce the number of constructs. This would also lead to fewer survey items especially considering that students need to provide answers for several scenarios.

To reduce the scope of the model, assumptions are made. AI-based assessment will be implemented in the higher education organisations irrespective of students' willingness to switch to such a system. Students are only informed of the new method of assessment before their open (essay) questions exams/assignments.

PEU construct is omitted since the context of the paper assumes that students would not have an interaction with the AI itself thus no effort would be needed to use the AI for the assessment.

SN construct is not included as well, although it is important to consider as per Sanchez-Prieto's work, in the context of this paper we assume that students do not have freedom of choice to use or not to use the AI assessment. They are simply presented with the fact that they will be assessed differently.

RC is not included based on the previous discussion about having no freedom of choice.

AU is defined as the frequency of use of AI assessment.

Keeping in mind the assumptions and that research should be done on the model itself (e.g., check if the model can explain the acceptance of the AI-based assessment) Figure 3 shows the conceptual framework used for this research.

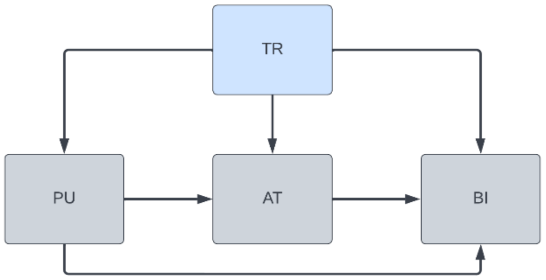


Figure 3. AI-based assessment acceptance (TAM AI)

The conceptual framework relevant to the aims of this research consists of PU, AT, BI and TR. The framework leaves the dependencies of the original untouched.

5.3 Hypothesis formulation

Before formulating the hypothesis, the application of the TAM AI will be explained. As was mentioned before, in relation to the research question, to understand how HitAI and its positioning will influence overall acceptance of the AI-based assessment of the students, the three scenarios will be presented:

- **Scenario 1 (S1):** no HitAI, the students will be assessed solely by the AI system, thus without any intervention by the teacher.

- **Scenario 2 (S2):** inclusion of HitAI, students will be assessed by the AI system, the teacher however will check the failed assignments to ensure that AI judgment was indeed correct (or incorrect). Students can also ask the teacher to review passed assignments.

- **Scenario 3 (S3):** inclusion of AI system solely as an advisory entity, the teacher has the last word in the assessment of student assignments.

A survey will be conducted to measure the constructs and statistical tests to establish the significance of the differences in answers when presented with different scenarios.

Additional tests should be done to establish the main drivers of the acceptance to understand the importance of constructs in acceptance of the AI-driven assessment.

Finally, students from different knowledge domains might have different levels of acceptance. It would be theorized that students from computer science will generally have higher acceptance levels.

The hypotheses are formulated as follows to aid the aims of the paper:

- **H1:** The acceptance of AI-driven assessment of students on the assignments with open questions will be higher in S3.

- **H2:** TR is the main driver of BI and thus has the most influence on the acceptance of AI-driven evaluation by students.

- **H3**: Respondents from the computer science study will have higher acceptance than all other respondents

6. RESEARCH DESIGN

This section explains how the survey was designed, how sampling was done and lastly how data was analysed.

6.1 Survey design

The design of the research will be adopted with three recommendations from (Sánchez-Prieto et al., 2020). It is done since the theory was adapted from the same paper.

First is to consider that the use of AI for the assessment is a technology that is in the early stages of development, and it needs recognition when measuring the constructs. Students can be confused if asked directly about AI. Thus, a brief introduction is presented before respondents start answering the questions. This is to avoid possible confusion about the nature of the survey.

The second is to recognise variables that can affect acceptance. These are identification variables such as gender, age, and the branch of knowledge. It will be outside of the scope to check for all such variables with relevant tests such as the MIMIC modelling used by (Teo et al., 2014). Thus, it is important to recognise the limitation of this study as it will not make use of such tools. However as defined in the H3, the branch of knowledge is considered.

Third, the Likert-type items will be used to measure the constructs. It is a common way to measure the constructs in TAM, some examples include: (Amoako-Gyampah, 2007), (Teo et al., 2014), (Teo et al., 2008) and (Castañeda et al., 2007). As the model was adopted and changed from Sanchez-Prieto's proposal, the items are altered to serve the aims of the research. One question will have five items ranging from 1 – strongly agree to 5 – strongly disagree. Survey items can be found in Appendix 3.

The survey includes the same questions for three different scenarios. Apart from an introductory explanation of AI assessment, the students will have a short description of scenarios so that they can give an answer based on three setups.

A quantitative approach was used via means of an online survey. Meaning that voluntary response sampling applies. The distribution of the survey was done via social media of the author, WhatsApp groups of study associations at the University of Twente and via SONA system available for researchers at the University of Twente. When students click the link in SONA and participate in the study, they are awarded 0.25 ECTS credits.

The distribution way means that some people will be more inclined to participate than others, it can perhaps attract a population that is more interested or has more experience with the AIs. That should be considered a limitation.

6.2 Sampling

The population that this research wants to study is the students that are currently being assessed by teachers. However, that is too broad and reaching students say from high school would be not feasible. Thus, students in higher education are the main aim of this study.

Subsequently, since the author has access to limited distribution networks, the sample will consist mostly of students in higher education at the University of Twente, Netherlands.

Another point to take is that this is non-probability sampling, leading to a chance of sampling bias. Non-response can be a problem as well as under coverage.

To reduce the risks of sampling bias, several steps were taken.

First, to ensure that responses are from students the survey includes the question that aims to clarify the current occupation.

Second, to ensure adequate coverage the survey was distributed in multiple chats with people from different branches of knowledge (groups from computer science, psychology and other).

Third, the survey will be as short as possible to be more accessible, and the introduction and scenarios will be short and clear as well. However, it can reduce the reliability of the answers, if the survey is long then positive/negative framing can be used for similar questions to check if the person's answer will be the same in both cases.

6.3 Data analysis

The choice of statistical tests depends on the type of Likert items. This research is using Likert scales since when the items are combined, they measure the acceptance of AI grading. Thus, the choice of tests is parametric, while using means and standard deviations to describe the scale (Boone et al., 2012).

This comes with assumptions about the data. Likert-type items cannot be normally distributed which is an assumption for using parametric tests. However, it is a common practice to analyse Likert scales in such a way, examples being (Teo et al., 2014), (Castañeda et al., 2007).

First descriptive statistics are presented for each of the scenarios to understand the general difference in answers. Statistical tests were then performed to check the hypothesises. First to find if there is a significant difference in answers between the three scenarios, then check for the difference in answers from participants in different study programmes.

Survey items were recoded as a means of answers. For instance, the survey consisted of 3 items for the PU variable, it was recoded as a single variable with execution code in SPSS as PU=MEAN (PU1, PU2, PU3). Since each question was asked 3 times about different grading scenarios/methods, the distinction is made between PU_AI, PU_Review, PU_Advisor. Variables stand for Scenario 1,2 and 3 respectively. Other variables were adjusted with the same principle.

The overall acceptance variables were recoded by combining all of the items resulting in, for instance, Acceptance_AI which means overall acceptance for scenario 1 with only an AI system and no human intervention.

Data analysis was completed using SPSS.

7. RESULTS

This section will present participants' statistics, then descriptive statistics and finally the testing of three hypotheses.

7.1 Participants

Participants were 63 students from (mostly) the University of Twente. Since the nature of the anonymous survey, it is not possible to ensure the location of participants when a link to the survey is shared online. However, it is known that 13 participants came from the SONA system internal to the University of Twente and links were shared in WhatsApp groups of the same university.

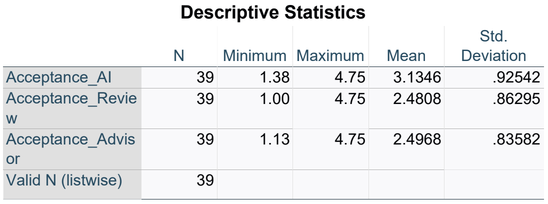
After data cleaning in SPSS, the number of respondents was intentionally reduced to 39. The reason is to exclude the non-completed surveys and responses with unrealistic completion times (under two minutes). The average completion time was 3.75 minutes (excluding artifactually large times due to respondents leaving the survey tab in the browser open and returning to it later).

48.7% were males, while 35.9% were female respondents. 15.4% account for other/prefer not to say options.

7.2 Descriptive statistics

First, it is important to remember that the scale used in this research differs from most other researchers. A 5-point scale was designed with 1 being strongly agree and 5 being strongly disagree. Thus, the means closer to 1 would indicate more positive answers.

The detailed descriptives by scenario can be found in Appendix 1. In this section, the overall acceptance is shown by combining all the survey items per scenario.



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Acceptance_AI	39	1.38	4.75	3.1346	.92542
Acceptance_Review	39	1.00	4.75	2.4808	.86295
Acceptance_Advisor	39	1.13	4.75	2.4968	.83582
Valid N (listwise)	39				

Table 1. Overall acceptance by scenario; Descriptive.

The means of all items are shown in Table 1. They are presented as the overall acceptance of three different scenarios.

1. The overall acceptance in scenario 1 (Acceptance_AI) has a **mean of 3.13** with an SD of 0.93.

2. The overall acceptance in scenario 2 (Acceptance_Review) has a **mean of 2.48** with an SD of 0.86.

3. The overall acceptance in scenario 3 (Acceptance_Advisor) has **a mean of 2.50** with an SD of 0.84.

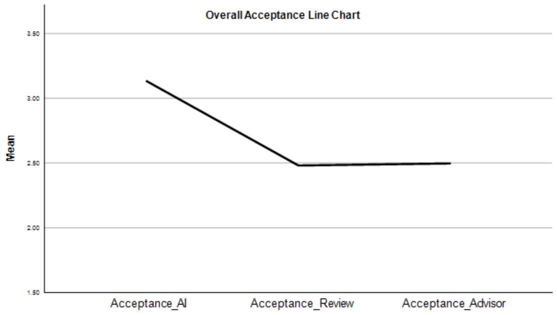


Figure 4. Overall acceptance mean chart. (The closer to 0 the more positive answers)

Figure 4 shows the difference in overall acceptance means visually, it is seen that acceptance with solely AI grading (scenario 1), greatly differs from two other scenarios.

Descriptive statistics indicate that scenarios two and three means do not differ too much and have higher overall acceptance than scenario one. Scenarios two and three include human-in-the-loop AI (HitAI) and are viewed as better alternatives to the AI-based grading system. If these differences are statistically significant or not is tested under hypothesis 1.

7.3 Hypothesis 1

The acceptance of AI-driven assessment of students on the assignments with open questions will be higher in S3.

ANOVA with Repeated Measures is used to test if there is a significant difference between means of acceptance in three different scenarios. This test was chosen since one respondent was asked to answer the same question three times with one difference, that is scenario. Overall acceptance means were used in this test.

First Mauchly's test, x2(2) =8.39, p=0.015 indicates a violation of sphericity. Meaning that sphericity is not assumed and thus in the test outcome, Greenhouse-Geisser correction is used.

Table 9 (Appendix 2) summarises the outcome of the test. Using the ANOVA with repeated measures with Greenhouse-Geisser correlation, the mean scores for acceptance of AI-driven evaluation were statistically significantly different (F (1.663, 63.177) = 13.139, p < 0.001).

Finally, post hoc analysis with a Bonferroni adjustment shows the mean differences in three scenarios. Scenario 1 without HitAI (without teacher intervention) is significantly different when compared to scenarios 2 and 3 (with HitAI), with p values being <0.001 and 0.002, respectively.

In contrast, scenario 2 where the teacher reviews some of the assignments and scenario 3 where the teacher grades every assignment while using AI as an advisor are not significantly different, with a p-value of 1. The mean difference between the two is 0.016.

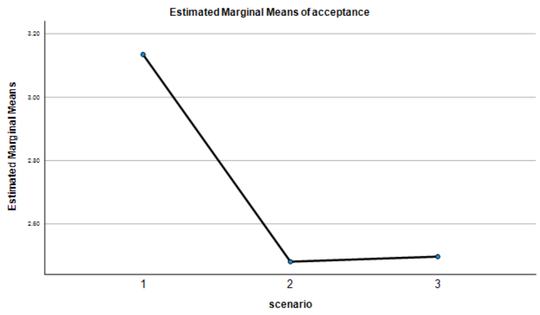


Figure 5. Difference in means of acceptance by scenario. (The closer to 0 the more positive answers)

Figure 5 provides better visualisation of differences between means of scenarios and summarises the last paragraph.

In conclusion, hypothesis 1 is not supported. From the graph and table, it is seen that the overall acceptance is higher in scenario 2, although not statistically significantly. That implies that overall acceptance is the same in both scenarios that include HitAI and thus human/teacher intervention. Acceptance of solely AI grading method is significantly lower compared to the other two.

Although the hypothesis is not supported, the author's initial thoughts were (more or less) correct in a way that system that includes humans would be accepted more than a fully AI system.

7.4 Hypothesis 2

TR is the main driver of BI and thus has the most influence on the acceptance of AI-driven evaluation by students.

A correlation matrix with Pearson correlation was used to determine the relationships between variables. The test was done three times for three different scenarios. This would allow determining the strongest correlation between all the variables in three different contexts.

All correlations in all three scenarios were statistically significant at the 0.01 level ($p < 0.01$). Remarkably, almost all relationships are strong. Ranging between 0.63 which is a moderate to 0.86 strong positive relationship, across all the scenarios.

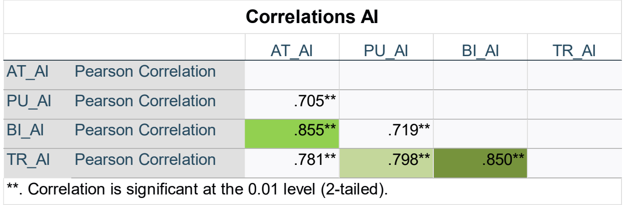
In the tables below you will find the outcomes of the tests for each of the scenarios. The bright green value highlights the strongest linear relationship, while the dark green cell shows the 2nd strongest relationship. Finally, greyed-out green displays 3rd strongest correlations, while white cells show the rest.

Table 2 shows the test outcome for scenario 1. Correlations are ranging from 0.705 to 0.855. The highest correlation is relevant here as it shows a relationship between BI and AT. The relationship between BI and TR has a coefficient of 0.850 which is not far off from the highest one. Both AT and TR can be the main drivers here. However, the initial hypothesis cannot be supported.

Table 3 displays the Pearson correlation for scenario 2. Correlations are ranging from 0.68 to 0.84. The second-largest coefficient is relevant in this case. TR is the main driver of BI in scenario 2. Thus, the hypothesis is supported. Remarkably, the relationship between AT and BI has a coefficient of 0.826 while the relationship between TR and BI is 0.827. It can be also concluded that they have the same strength of the relationship.

Finally, Table 4 shows the coefficients in scenario 3. They are ranging from 0.63 to 0.86. Same as in the first scenario highest coefficient is relevant. In this case, again, the AT is the main driver of BI. The relationship between TR and BI, in this case, has a coefficient of 0.79.

In conclusion, hypothesis 2 cannot be supported for every scenario. It is only true for scenario 2. However, in other scenarios, the AT is the main driver for BI and so it has the most influence on the acceptance of AI-driven evaluation by students. Important to note, that TR has almost the same strength of the relationship as AT in scenarios 1 and 2.

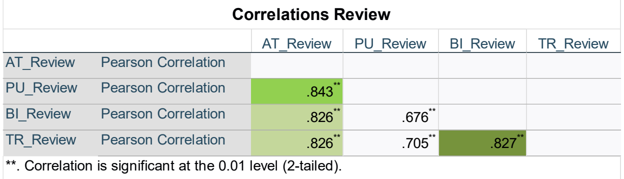


Correlations AI

		AT_AI	PU_AI	BI_AI	TR_AI
AT_AI	Pearson Correlation				
PU_AI	Pearson Correlation	.705**			
BI_AI	Pearson Correlation	.855**	.719**		
TR_AI	Pearson Correlation	.781**	.798**	.850**	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 2. Scenario 1. Pearson correlation matrix.

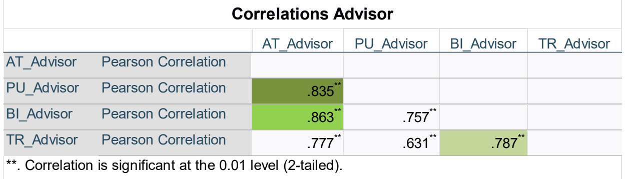


Correlations Review

		AT_Review	PU_Review	BI_Review	TR_Review
AT_Review	Pearson Correlation				
PU_Review	Pearson Correlation	.843**			
BI_Review	Pearson Correlation	.826**	.676**		
TR_Review	Pearson Correlation	.826**	.705**	.827**	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 3. Scenario 2. Pearson correlation matrix.



Correlations Advisor

		AT_Advisor	PU_Advisor	BI_Advisor	TR_Advisor
AT_Advisor	Pearson Correlation				
PU_Advisor	Pearson Correlation	.835**			
BI_Advisor	Pearson Correlation	.863**	.757**		
TR_Advisor	Pearson Correlation	.777**	.631**	.787**	

**. Correlation is significant at the 0.01 level (2-tailed).

Table 4. Scenario 3. Pearson correlation matrix.

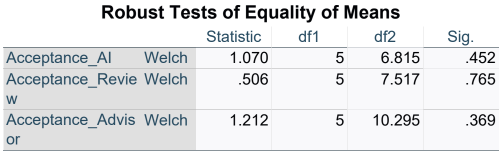
7.5 Hypothesis 3

Respondents from the computer science study will have higher acceptance than all other respondents.

Usually, one-way ANOVA would be used to check for the difference in means of the field of study. However, one of the assumptions is violated when trying to conduct the test, namely the assumption of equal variances.

Thus, the Welch test was used, it assumes that variances are not equal.

Before conducting this test, four observations were filtered. The Welch test cannot be performed when one group has less than two cases. The survey had only one respondent per medicine, law, chemistry, and biology studies. Therefore, they were filtered out leaving business, computer science, psychology, sociology, arts, and other studies.



Robust Tests of Equality of Means

		Statistic	df1	df2	Sig.
Acceptance_AI	Welch	1.070	5	6.815	.452
Acceptance_Review	Welch	.506	5	7.517	.765
Acceptance_Advisor	Welch	1.212	5	10.295	.369

The results are the following:

1) Overall mean acceptance of AI in scenario 1 does not differ significantly across study fields, F (5,6.82) = 1.07, p = 0.452

2) Overall mean acceptance of AI in scenario 2 does not differ significantly across study fields, F (5,7.52) = 0.51, p = 0.765

3) Overall mean acceptance of AI in scenario 3 does not differ significantly across study fields, F (5,10.30) = 1.21, p = 0.369

None of the means were significantly different which indicates that hypothesis 3 is not supported. It was hypothesised based on the literature review that students from computer science might have higher overall acceptance when in turn it is not true. There is no difference in means across the study fields.

8. DISCUSSION AND CONCLUSION

The research aimed to find out the acceptance levels of AI-driven evaluation of essay questions in higher education from the perspective of students. Then finding out what approach will yield the highest acceptance levels.

The results indicate that approaches to AI evaluation with human intervention (HitAI) yield higher acceptance than the AI system without HitAI. This study used two approaches/scenarios for an AI-driven grading system with human intervention. Both approaches score the same on the acceptance level.

Identical acceptance levels between scenarios with HitAI were unexpected. That could indicate that students do not want to be graded only with the AI and would prefer some intervention from the teacher, and it does not matter to what degree the teacher is present in the system. As long as there is a way for a teacher to look at the assignment and grade it manually. That would be of high importance to the universities that are planning on implementing such a system. Students would need to know that there is still a connection with the previous "teacher" system.

As the topic of this study represents the underexplored area of research it is problematic to find any evidence that supports or contradicts the results from other papers. A study by (Marcinkowski et al., 2020) about using AI in university admissions found that AI is perceived as much fairer agent compared to a human committee. Which was in line with other studies in the field stating that high-impact decisions when done by AI are perceived as fairer compared to humans. Assuming that getting a final grade for the exam is a high-impact decision, one would believe that acceptance should be higher for the AI system and not systems with HitAI. Which is not the case. But of course, this research focus is not on fairness and outcomes of the AI and human systems were not compared.

Furthermore, it was hypothesized that trust will have the most influence on acceptance. This hypothesis is not supported. In two scenarios the attitudes toward the use of AI have the most influence over the acceptance of AI. While in one scenario, trust has the most influence on acceptance.

In scenario 1 (no HitAI) the acceptance of AI-driven grading is influenced highly by attitudes towards AI use and trust.

In scenario 2 (HitAI) the main drivers of the acceptance are attitudes and trust as well.

In scenario 3 (HitAI) the acceptance is strongly influenced by attitudes.

In all three scenarios, the attitudes toward AI use are the main driver for the AI-driven grading acceptance. Attitudes describe the feelings and assessments of the AI systems. Meaning, that the students think more in irrational terms when it comes to their assessment by the AI. Perceived usefulness, which can be associated more with rational thinking about the technology, has lower coefficients across the three scenarios.

Interestingly, in scenario 3 the acceptance is (highly) influenced only by attitudes and not trust. Perhaps, in the 3rd scenario with high teacher involvement, trust becomes less of a concern as the assessment method does not differ greatly from traditional teachers grading.

In contrast, in scenarios 1 and 2 where the AI involvement in the grading process is higher, trust becomes more relevant in the eyes of students. Questions concerning the trust and accuracy of evaluation would strongly influence the acceptance of such AI evaluation. This is in line with the initial hypothesis. However, the effect of attitudes was not expected. Additionally, it is problematic to compare the findings with other research as there are no scientific papers on the subject. Perhaps future research could investigate the implications of these results.

Moreover, it was theorized that students in computer science would have higher acceptance levels. The reasoning is that computer science students would have better knowledge of AI. However, the results indicate that field of study does not matter when it comes to the overall acceptance. This finding is contradicting research by (Marcinkowski et al., 2020) on using AI systems in university admissions. Their findings suggest that there should be a difference in the acceptance of AI-based systems in the field of study. In their research students from mathematics and natural sciences show less intention to protest against the AI system and do not show exit behaviours. However, it is important to mention that the sample size in this research was considerably smaller than the sample in the paper on admissions. Additionally, this research sample did not include any students from natural sciences or mathematics fields.

The thoughts on findings are somewhat speculative and more research should be done on the topic to start a discussion.

8.1 Limitations

It is important to talk about the limitations of this study. They include possible sampling bias and the absence of rigorous testing of differences between groups using MIMIC modelling for instance.

Additionally, it should be noticed that the proposed model was not evaluated as was done for instance by Teo et al., (2014). In their research, the modified TAM model was assessed using confirmatory factor analysis and estimated using the maximum likelihood estimation procedure. The results would indicate if items were reliable indicators of the hypothesized constructs they were supposed to measure. However, this research does not include such analysis and should be considered a limitation.

Additionally, the research model is largely based on the proposal by Sánchez-Prieto et al., and no research was done with their model. However, the authors of the proposed model aim to test the model they proposed and conduct research in the future.

Next, the possible explanation of no difference in overall acceptance in both scenarios with HitAI could be explained by misunderstanding by participants on how exactly methods differ.

Lastly, the sample mostly consists of students in one Dutch university. Repeated research with participants in other universities could give different results. Furthermore, the sample size is small, consisting of only 39 participants. The size of the sample is not necessarily small however larger sample consisting of students from different universities might contradict the findings of this paper.

8.2 Future Research

Future studies can address these limitations. It would be possible to repeat this study with MIMIC modelling and alike. Additionally, the tests can be done on model fit. Finally, research may be conducted in other geographical areas with larger sample sizes.

While this paper's findings conclude that human intervention in AI indicates overall higher acceptance, this, in turn, raises the question of the different methods of keeping humans in the loop that influences the overall acceptance. Perhaps in this study, both scenarios with teacher intervention were too similar. In a way in both scenarios teacher still can grade the assignments. Only in one scenario students would need to go the extra step to ask the teacher to review the work. This calls for a new study with more methods for keeping humans in the loop. As the found no difference between both HitAI scenarios might be a limitation.

8.3 Conclusion

In general, this research provides a view into student perspectives of AI in higher education. As we know universities might want to adopt AI-driven evaluations for essay questions for multiple reasons such as being more efficient at grading. However, when done incorrectly students might raise their voices against using such systems which in turn will influence the university's reputation. Acceptance of the grading process should be an important aspect to consider for the university.

Based on findings in this paper it is suggested for universities to adopt AI grading methods while keeping teachers in the loop. Findings also suggest that there is no difference in acceptance between different study fields. Higher education institutions should consider students' trust in the system and their attitudes toward the AI-based assessment. This is because trust and students' attitudes have the strongest influence on acceptance.

Altogether, students' acceptance is important to consider when planning to implement the AI-driven evaluations of essay questions in the university.

9. REFERENCES

Amoako-Gyampah, K. (2007). Perceived usefulness, user involvement and behavioral intention: an empirical study of ERP implementation. Computers in Human Behavior, 23(3), 1232–1248. https://doi.org/10.1016/J.CHB.2004.12.002

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI and Society, 35(3), 611–623. https://doi.org/10.1007/S00146-019-00931-W/TABLES/4

Boone, H. N., Associate Professor, J., & Boone Associate Professor, D. A. (2012). Number 2 Article Number 2TOT2 (Vol. 50). http://www.joe.org/joe/2012april/tt2p.shtml[8/20/20129:07:48AM]

Castañeda, J. A., Muñoz-Leiva, F., & Luque, T. (2007). Web Acceptance Model (WAM): Moderating effects of user experience. Information & Management, 44(4), 384–396. https://doi.org/10.1016/J.IM.2007.02.003

Cranor, L. F. (2008). A Framework for Reasoning About the Human in the Loop.

Cruz-Benito, J., Sánchez-Prieto, J. C., Therón, R., & García-Peñalvo, F. J. (2019). Measuring Students' Acceptance to AI-Driven Assessment in eLearning: Proposing a First TAM-Based Research Model. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11590 LNCS, 15–25. https://doi.org/10.1007/978-3-030-21814-0_2/FIGURES/3

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly: Management Information Systems, 13(3), 319–339. https://doi.org/10.2307/249008

Gierl, M. J., Latifi, S., Lai, H., Boulais, A. P., & de Champlain, A. (2014). Automated essay scoring and the future of educational assessment in medical education. Medical Education, 48(10), 950–962. https://doi.org/10.1111/medu.12517

Haton, J. P. (2006). A brief introduction to artificial intelligence. IFAC Proceedings Volumes, 39(4), 8–16. https://doi.org/10.3182/20060522-3-FR-2904.00003

Lilly, J. D., Wipawayangkool, K., & Pass, M. (2022). Teaching Evaluations and Student Grades: That's Not Fair!: Https://Doi-Org.Ezproxy2.Utwente.Nl/10.1177/10525629221084338. https://doi.org/10.1177/10525629221084338

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (Un-)Fairness in Higher Education Admissions The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation. https://doi.org/10.1145/3351095.3372867

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. Artificial Intelligence Review, 55(3), 2495–2527. https://doi.org/10.1007/S10462-021-10068-2/TABLES/9

Sánchez-Prieto, J. C., Cruz-Benito, J., Therón, R., & García-Peñalvo, F. (2020). Assessed by Machines: Development of a TAM-Based Tool to Measure AI-based Assessment Acceptance Among Students. International Journal of Interactive Multimedia and Artificial Intelligence, 6(4), 80. https://doi.org/10.9781/ijimai.2020.11.009

Teo, T., Ruangrit, N., Khlaisang, J., Thammetar, T., & Sunphakitjumnong, K. (2014). EXPLORING E-LEARNING ACCEPTANCE AMONG UNIVERSITY STUDENTS IN THAILAND: A NATIONAL SURVEY. J. EDUCATIONAL COMPUTING RESEARCH, 50(4), 489–506. https://doi.org/10.2190/EC.50.4.c

Teo, T., Su Luan, W., & Sing, C. C. (2008). A cross-cultural examination of the intention to use technology between Singaporean and Malaysian pre-service teachers: an application of the Technology Acceptance Model (TAM). Educational Technology & Society, 11(4), 1176–3647.

TURING, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/MIND/LIX.236.433

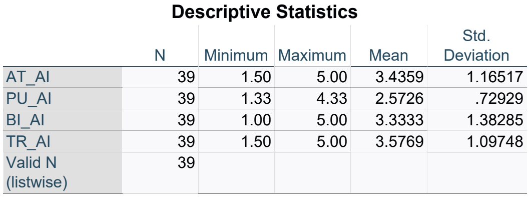
Terzis, V., & Economides, A. A. (2011). The acceptance and use of computer based assessment. Computers & Education, 56(4), 1032–1044. https://doi.org/10.1016/J.COMPEDU.2010.11.017

Zanzotto, F. M. (2019). Viewpoint: Human-in-the-loop Artificial Intelligence. Journal of Artificial Intelligence Research, 64, 243–252. https://doi.org/10.1613/JAIR.1.11345

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? International Journal of Educational Technology in Higher Education 2019 16:1, 16(1), 1–27. https://doi.org/10.1186/S41239-019-0171-0

10. APPENDICES

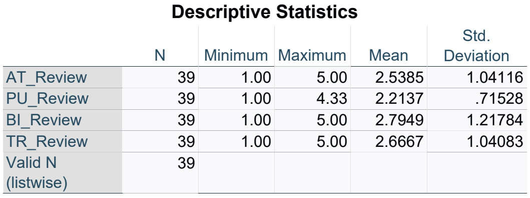
10.1 Appendix 1: Descriptives by scenario



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AT_AI	39	1.50	5.00	3.4359	1.16517
PU_AI	39	1.33	4.33	2.5726	.72929
BI_AI	39	1.00	5.00	3.3333	1.38285
TR_AI	39	1.50	5.00	3.5769	1.09748
Valid N (listwise)	39				

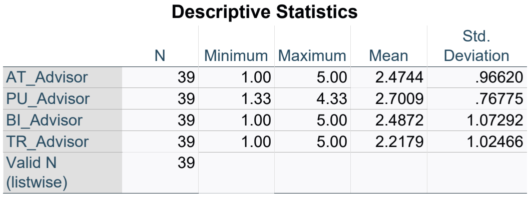
Table 5. Scenario 1: Students are assessed solely by the AI.



Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AT_Review	39	1.00	5.00	2.5385	1.04116
PU_Review	39	1.00	4.33	2.2137	.71528
BI_Review	39	1.00	5.00	2.7949	1.21784
TR_Review	39	1.00	5.00	2.6667	1.04083
Valid N (listwise)	39				

Table 6. Scenario 2: Students are assessed by the AI; teachers review failed assignments and review can be requested by the students.

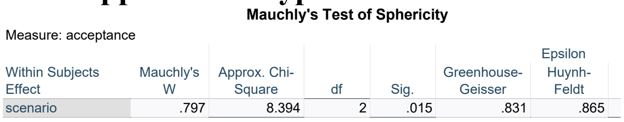


Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
AT_Advisor	39	1.00	5.00	2.4744	.96620
PU_Advisor	39	1.33	4.33	2.7009	.76775
BI_Advisor	39	1.00	5.00	2.4872	1.07292
TR_Advisor	39	1.00	5.00	2.2179	1.02466
Valid N (listwise)	39				

Table 7. Scenario 3: The last grading decision is given to the teacher, AI as an advisor.

10.2 Appendix 2: Hypothesis 3 tests

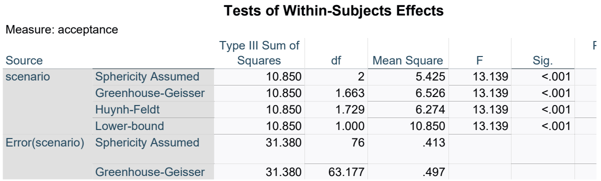


Mauchly's Test of Sphericity

Measure: acceptance

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon	
					Greenhouse-Geisser	Huynh-Feldt
scenario	.797	8.394	2	.015	.831	.865

Table 8. Test of sphericity.

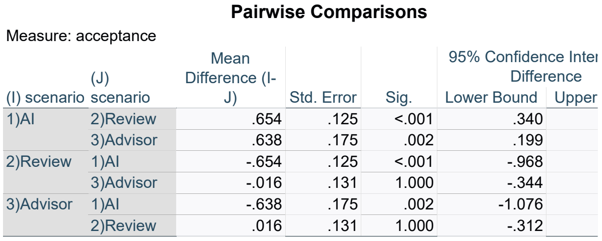


Tests of Within-Subjects Effects

Measure: acceptance

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	F
scenario	Sphericity Assumed	10.850	2	5.425	13.139	<.001	
	Greenhouse-Geisser	10.850	1.663	6.526	13.139	<.001	
	Huynh-Feldt	10.850	1.729	6.274	13.139	<.001	
	Lower-bound	10.850	1.000	10.850	13.139	<.001	
Error(scenario)	Sphericity Assumed	31.380	76	.413			
	Greenhouse-Geisser	31.380	63.177	.497			

Table 9. Greenhouse-Geisser correlation.



Pairwise Comparisons

Measure: acceptance

(I) scenario	(J) scenario	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Inter	
					Difference	
					Lower Bound	Upper
1)AI	2)Review	.654	.125	<.001	.340	
	3)Advisor	.638	.175	.002	.199	
2)Review	1)AI	-.654	.125	<.001	-.968	
	3)Advisor	-.016	.131	1.000	-.344	
3)Advisor	1)AI	-.638	.175	.002	-1.076	
	2)Review	.016	.131	1.000	-.312	

Table 10. Post hoc analysis.

10.3 Appendix 3: Survey items

Perceived Usefulness (PU)

PU1: Using these AI grading methods would be useful.

PU2: Using these AI grading methods will benefit my academic productivity (for instance, exam results will be available quicker).

PU3: Using these AI grading methods increases my assessment opportunities.

Attitudes towards use (AT)

AT1: It is a good idea to implement these AI grading methods for open questions exams or assignments.

AT2: I have positive feelings about implementing these AI grading methods for open questions exams or assignments.

Behavioural intention (BI)

BI1: I hope my university will implement these AI grading methods for open questions exams or assignments.

Trust (TR)

TR1: I trust these AI grading systems.

TR2: I believe these AI grading systems are accurate.