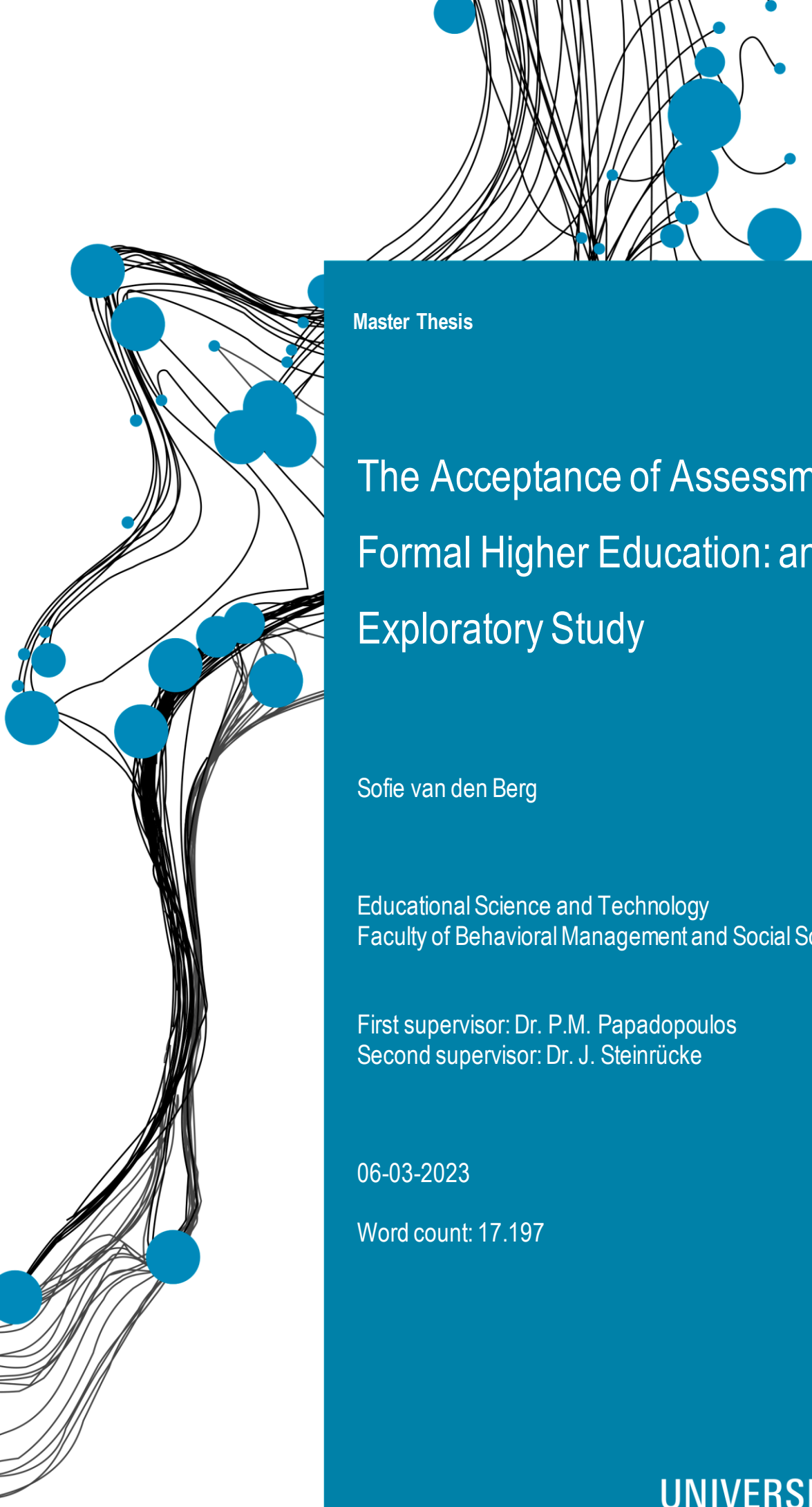

Master Thesis

The Acceptance of Assessment AIs in Formal Higher Education: an Exploratory Study

Sofie van den Berg

Educational Science and Technology
Faculty of Behavioral Management and Social Sciences

First supervisor: Dr. P.M. Papadopoulos
Second supervisor: Dr. J. Steinrücke

06-03-2023

Word count: 17.197

UNIVERSITY OF TWENTE.

Abstract

Background There has been an increase in research into the usage of Artificial Intelligence (AI) in education. AI has been increasingly incorporated into the educational system, bringing promises of efficiency and effectiveness. One of those areas is assessment. However, little research is dedicated to the attitudes and technology preferences of two important stakeholders: students and teachers. Using the theoretical models of UTAUT and the Technology Readiness Index, a framework was created through which the acceptance of AI assessment could be measured. An exploratory approach was taken to gain insights into the personal attitudes and technology preferences of students and teachers at the University of Twente.

Methods This research consisted of two phases: a survey and interviews. The survey consisted of a baseline technology acceptance questionnaire and a set of scenarios. The scenarios provided participants with hypothetical situations in which AI assessment was used in an educational context. Data from 20 participants was gathered (12 students and 8 teachers). Both quantitative and qualitative data was gathered through the survey instrument. The interviews were based on participants' responses and were held with 12 participants (7 students and 5 teachers). Qualitative analysis was performed on the interviews, using in-vivo coding.

Results Survey results showed that both participant groups scored high on Performance Expectancy. This was mirrored in scenario responses, where participants favored scenarios that benefit their personal gain. There was no clear relation between baseline scores and scenario acceptance in both groups. For students, 13 themes were found in the analysis of the interviews. For teachers, 10 themes were found in the analysis of the interviews. For both groups, Perceived Risk and Performance Expectancy were predominantly present in the interviews.

Conclusion Both groups suffer from inexperience with AI technology and therefore have low trust in its assessment. Performance Expectancy was found to be an important driver for AI assessment acceptance. The following guidelines were suggested based on the findings: students and teachers need knowledge training and positive experiences with AI assessment before implementation, AI assessment should start its implementation in formative assessment settings and only provide assisting feedback in summative settings, and full program transparency should be offered to both students and teachers.

Contents

List of Figures

List of Tables

1 Introduction

There has been an increase in research into Artificial Intelligence (AI). Cugurullo (2020) defined AI as "the integration of artificial (not a natural process, but one induced by machines) and intelligence (skills of learning, to extract concepts from data and to handle uncertainty in complex situations)". This type of technology is characterized by simulating human intelligence activities and adopting behaviors like learning, judgment, and decision-making (Xu, Lu, & Li, 2021).

AI has been recognized as an important technological development in educational technology, with researchers linking the technology to the future of education (Holmes et al., 2021). The technology has been used for various applications, e.g., chatbots, learning analytics, and intelligent tutors (Zawacki-Richter, Marín, Bond, & Gouverneur, 2019). While this type of technology provides multiple benefits in our current educational system (i.e., reduced workload for teachers (Owoc, Sawicka, & Weichbroth, 2021), personalized learning plans for students (Bajaj & Sharma, 2018), and continuous assessment (Holmes et al., 2021)), it also brings new risks and concerns (Borenstein & Howard, 2021). Especially in education, the lack of guidelines, policies, and regulations makes for a difficult position (du Boulay, Poulovasillis, Holmes, & Mavrikis, 2018). Researchers highlighted a gap in knowledge on reflection on AI usage, which is necessary to successfully implement AI in education (Holmes et al., 2021).

Currently, limited research has been conducted into what students and teachers want to see in AI applications (Holmes et al., 2021). Jaakkola, Henno, Lahti, Jarvinen, and Makela (2020) highlighted in their research that the integration of AI in lesson programs comes with drastic changes for both teacher and student. These changes include the need for schooling in technologies and a re-division of teachers' workload. Often, studies analyzing the use of AI in education focus on the technical approach (i.e., does the program do what it is supposed to do) instead of the underlying pedagogical or ethical reasoning (González-Calatayud, Prendes-Espinosa, Roig-Vila, & Carpanzano, 2021; Holmes et al., 2021). González-Calatayud et al. (2021) called for a collaboration between AI experts and educational experts to simultaneously understand the users' needs as well as the potential of the technology. Currently, there is a knowledge gap in how the users of AI will react to its services and to what extent this affects their technology acceptance. Both fields need to work together in order to fully grasp the benefits and challenges of AIED.

Since the range of different AI applications is so broad, the focus of this thesis will be on AI approaches that grade and assess students' assignments since Cope, Kalantzis, and Searsmith (2021) claimed that assessment might be the area in education with the best opportunity to implement AI. This type of AI grades students based on the performed task (e.g., essays, language proficiency, and exams) (González-Calatayud et al., 2021).

In the following parts, a more excessive background is given on the use of AI in the educational assessment. To research this concept and the acceptance thereof, an adapted model of the Unified Theory of Acceptance and Use of Technology (UTAUT) and the Technology Readiness Index (TR) will be presented and used for this thesis.

1.1 AI Assessment in Education

Assessment is essential to determine and gauge the learning process of students (Swiecki et al., 2022). The most common types of assessment in Higher Education are formative and summative assessment. Formative assessment takes place during the run of a course to check whether the

student is making learning progress, and summative assessment takes place at the end to verify learning outcomes (Harlen & James, 1997). In both areas, assessment tasks can be prone to errors and seen as tedious by teachers (Vittorini, Menini, & Tonelli, 2021). Traditional assessment provides snapshots of learning outcomes, while AI can transform this into incremental and adaptive learning pathways.

An important development in AI assessment is Automated Essay Scoring (AES). AES is an assessment system that bases student essay scores on a list of textual features (Uto, 2021). These features can include: word frequency, academic language, and contextual distinctiveness (Gardner, O'Leary, & Yuan, 2021). These models create a score based on the aforementioned features, using human scoring decisions as input for their decisions (Uto & Okano, 2021; Ferrara & Qunbar, 2022). In a way, educational assessment and machine learning apply the same underlying concepts. The program needs to be 'taught' the content of the assessment, then it can assess students whether they applied their knowledge adequately. When the program is capable of understanding quality criteria in students' responses, it is capable of assessing content (Somasundaran, Lee, Chodorow, & Wang, 2015). The future of AES seems bright, but its disadvantages should be taken into account as well.

Promises of AES include time-saving for teachers, improving consistency, and being cost-effective (Uto & Okano, 2021; Beseiso, Alzubi, & Rashaideh, 2021). In addition, it also promises the elimination of human bias (i.e., fatigue, influence of the assessor's expertise, inconsistency) (Taghipour, 2017). Although, the complete removal of bias seems to be an overestimated effect that AES systems can have, as they might suffer from an inconsistent training data pool or rely on spurious correlations. In fact, there is a growing concern that the automation of assessment might increase the unfairness rate: studies show that e-raters sometimes tend to give higher scores to students from specific demographics as compared to human raters (Bridgeman, Trapani, & Attali, 2009; Litman, Zhang, Correnti, Matsumura, & Wang, 2021). Opposite, GUO (2009) found no trace of unfair treatment when comparing humans and e-raters. It is important to note that these studies focus on comparing assessment standards with human raters, who are not free from bias themselves either (Doewes, Saxena, Pei, & Pechenizkiy, 2022; Litman et al., 2021).

Another downside is that AES has rarely been developed on a rubric level and often operates on a black-box concept. Its biggest issue rests on a lack of transparency and explainability (Kumar & Boulanger, 2021). The transparency of a system entails whether the process can be described, whereas the explainability of a system entails whether it can be explained how and why a system arrived at a certain score (Arrieta et al., 2020; Roscher, Bohn, Duarte, & Garcke, 2020). The output of AES systems is often a holistic grade, without explaining how the system has come to that grade (Ke & Ng, 2019; Kumar & Boulanger, 2020). This is problematic since students need more than an overall grade that represents the quality of their work. Recent research has been dedicated to designing rubrics that allow AES systems to grade parts of assignments in isolation (Taghipour, 2017). These dimensions scores can give students the necessary information on which areas of their assignments have room for improvement (Ke & Ng, 2019). However, Powers, Burstein, Chodorow, Fowles, and Kukich (2002) concluded that AES systems could be easily fooled whenever their parameters for good scores were known and thus were not ready to take over the task of the sole scorer.

As Kumar and Boulanger (2020) pointed out in their review, fully automating assessment might not be possible in the near future. Certain aspects of assessment rubrics (e.g., ideas and concepts) remain out of AI's data set's reach. They propose a human-AI fusion when using AES systems to still decrease teachers' workload as well as acquire data to better train these systems. Above all,

there should be an organizational capability before any AI service can be implemented. Machicao (2019) expressed that universities should work on the human capacity to embrace AI services and divide more resources into increasing trust from all actors involved. This calls for research into how people decide whether to adopt technology.

1.2 Theoretical Framework

In this section, the theoretical framework of this thesis is presented. To measure the level of technology acceptance, the Unified Theory of Acceptance and Use of Technology (UTAUT) was chosen (Venkatesh, Morris, Davis, & Davis, 2003). In addition, the Technology Readiness Index (TR) was used to measure the personal attitude of users toward technology in general (Parasuraman, 2000). The two models are explored below and the resulting framework is presented concludingly.

1.2.1 The Unified Theory of Acceptance and Use of Technology

Technology acceptance models are derived from cognitive models and focus on the concept that individuals develop certain attitudes toward objects. These attitudes, in turn, define whether a user has the intention to perform consistent adoption behavior. The Unified Theory of Acceptance and Use of Technology (UTAUT) by Venkatesh et al. (2003) is a well-known technology acceptance model. The UTAUT model includes constructs of eight prior models, including the theories of Planned Behavior and Technology Acceptance Model (Chang, 2012). This theory has effectively contributed to research in technology acceptance (Chatterjee, Rana, Khorana, Mikalef, & Sharma, 2021). Raffaghelli, Rodríguez, Guerrero-Roldán, and Bañeres (2022) highlighted that the usage of the UTAUT model, especially in the field of AI, can provide new and needed insights into the acceptance of this technology.

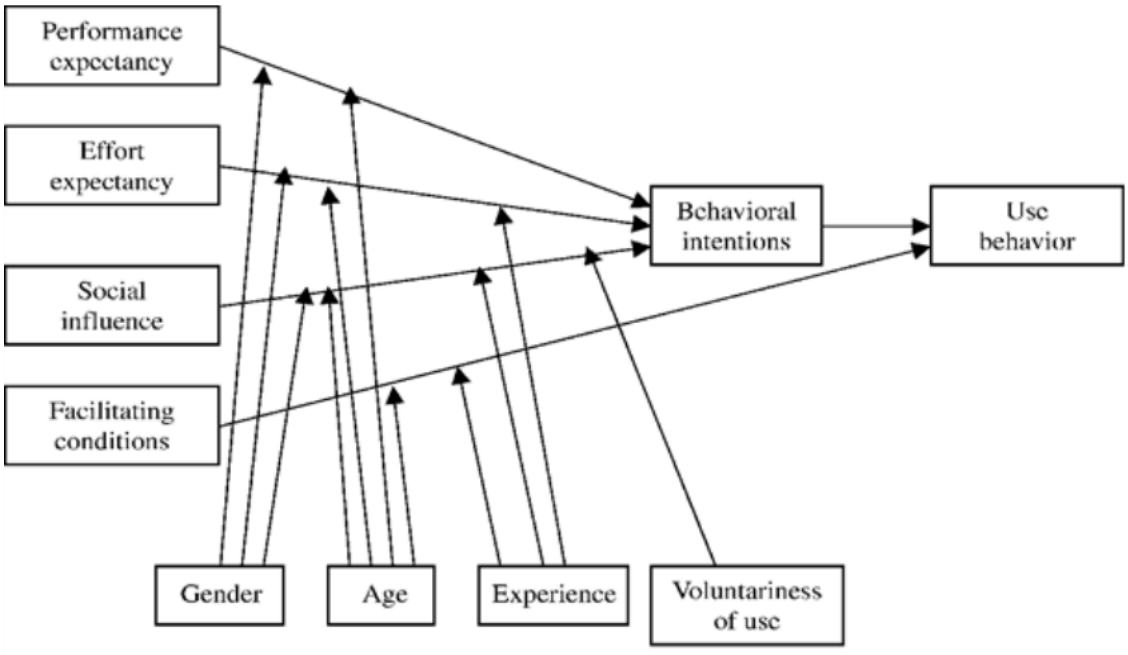

Figure 1. Unified Theory of Acceptance and Use of Technology by Venkatesh et al. (2003).

Originally, the UTAUT model consists of four core variables, accompanied by four mediating variables, see Figure 1. These four variables all influence a user's behavioral intention to use a specific technology, which in turn, affects the actual usage behavior (Venkatesh et al., 2003). The variables of interest for the scope of this thesis are Performance Expectancy and Effort Expectancy, whereas Chatterjee and Bhattacharjee (2020) also included Perceived Risk in their conceptual model.

Performance Expectancy (PE) refers to the extent to which an individual sees added value in the usage of a certain technology (Venkatesh et al., 2003). In the context of this study, PE is the belief that AI brings added value to the concept of assessment in education and will bring satisfactory benefits. Fridin and Belokopytov (2014) and Andrews, Ward, and Yoon (2021) have both shown that PE is a strong predictor of the intention of technology adoption.

Effort Expectancy (EE) refers to the extent to which an individual perceives the interaction with a certain technology as effortsome, much like perceived ease of use (Venkatesh et al., 2003). In the setting of this study, EE is seen as how easy the interaction is between an AI system and its users. While EE has not been shown to be a great predictor of intention to adopt as compared to PE, research shows that this construct is still of significance in the adoption process (Andrews et al., 2021; Aharony, 2015).

Perceived Risk (PR) refers to the perceived loss an individual would experience when using a certain technology (Warkentin, Gefen, Pavlou, & Rose, 2002). The presence of PR indicates that users experience disadvantages when being confronted with AI programs in assessment. Teo and Liu (2007) have found that perceived risk negatively affects an individual's attitude toward technology.

The UTAUT model has been used to predict whether certain populations are willing to adopt AI in their work routine. Handoko and Liusman (2021) assessed whether external auditors were willing to adopt an AI algorithm to aid in fraud detection. Their conclusion was that perceived usefulness and relative advantage were the most significant factors in the adoption choice of external auditors. Similarly, Chatterjee and Bhattacharjee (2020) found an 84 percent explanation rate for their version of the UTAUT model on the acceptance rate of AI in Indian higher education. However, they conclude that, at the time of conducting their study, AI adoption was at a crawling speed in India. This could have implications for the explainability of their model. The UTAUT model is used to measure the acceptance of technological characteristics, making it more context-specific. However, it is essential to point out that technology acceptance is not only affected by those factors, but also by personal attributes.

1.2.2 Technology Readiness Index

Additionally, the Technology Readiness Index measures the mental state of a consumer and determines to what extent they are willing to adopt a certain new technology. Parasuraman (2000) defined Technology Readiness (TR) as follows: "Technology readiness represents a gestalt of mental motivators and inhibitors that collectively determine a person's predisposition to use new technologies". This mental state consists of both positive and negative aspects which, combined, lead to the decision whether to adopt or not (M. L. Lai, 2008).

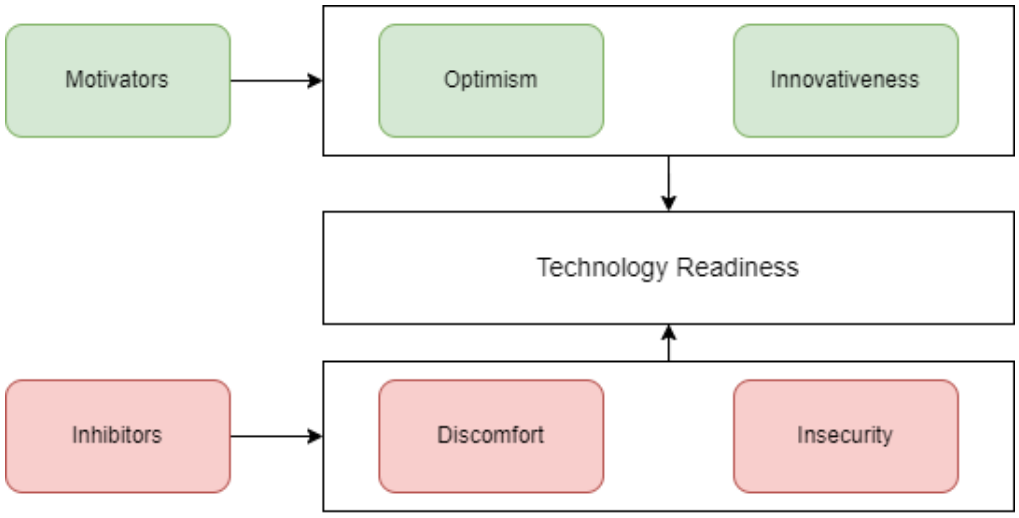

Figure 2. Technology Readiness Index by Parasuraman (2000).

Within the TR construct, two variables are defined: Motivators and Inhibitors, as shown in Figure 2.

Motivators are the positive aspects of TR and include the traits of innovativeness and optimism. Optimism refers to the thought that technology is a good thing in the individual's life as seen in factors such as an increased sense of control, flexibility, and efficiency. (Parasuraman, 2000; Tsikriktsis, 2004). Innovativeness refers to innovative behavior regarding adopting new technologies (Parasuraman, 2000). Individuals who are innovative are keen to try out features and updates of new technology and are highly motivated to accept novel inventions (Kuo, Liu, & Ma, 2013). Especially optimism has been found to have positive influences on technology adoption in different contexts (Kuo et al., 2013).

Inhibitors are the negative aspects of TR and include the traits of discomfort and insecurity. Discomfort refers to the feeling of being overwhelmed and having a lack of control over new technologies. Insecurity refers to distrust and skepticism toward technology and an overall feeling of negative consequences of its usage (Parasuraman, 2000). Individuals with low levels of security have less confidence in technology (Parameswaran, Kishore, & Li, 2015). Rahman, Taghizadeh, Ramayah, and Alam (2017) found that insecure agents thought that using a certain technology too often would decrease the quality of interpersonal relationships with their clients.

The TR scale has been used to measure technology preferences within a certain population to assess whether there will be implications when implementing a certain technology (Parasuraman & Colby, 2015). Damerji and Salimi (2021) used the TR construct in their research on AI adoption among accounting students. While their results did shine a positive light on the effects of TR on technology adoption, they recommend exploring other factors and constructing relationships alongside TR to ensure a complete picture. In addition, Blut and Wang (2020) highlighted that it is necessary to explore other mediators besides TR to explain the attitude-behavior gap.

1.2.3 Current Theoretical Framework

Figure 3 shows the composed theoretical framework for the purpose of this study. As mentioned before, the constructs of Perceived Risk, Performance Expectancy, and Effort Expectancy have been taken from the UTAUT model by Venkatesh et al. (2003) and the constructs of Inhibitors and Motivators have been taken from the Technology Readiness Index of Parasuraman (2000). This is where TR adds value to the framework, as it focuses on the personal characteristics of the user (Rinjany, 2020).

A combination of TR and other technology acceptance models is not a novel concept. Kuo et al. (2013) combined TR with TAM to assess the acceptance of electronic record systems within the nursing population. In a similar way, Y. L. Lai and Lee (2020) theorized that users' TR perception influences their intention to use a specific technology. They found that the variables optimism and innovativeness positively correlated with the constructs of usefulness and ease of use and that optimism significantly impacted attitude. As well as Rahman et al. (2017), in which they used TR in combination with the constructs of PE and EE of the UTAUT model to assess technology adoption in micro-entrepreneurs. Similarly, Cabrera-Sánchez, Villarejo-Ramos, Liébana-Cabanillas, and Shaikh (2021) extended the UTAUT model with the variables technology fear and consumer trust, closely representing the TR model.

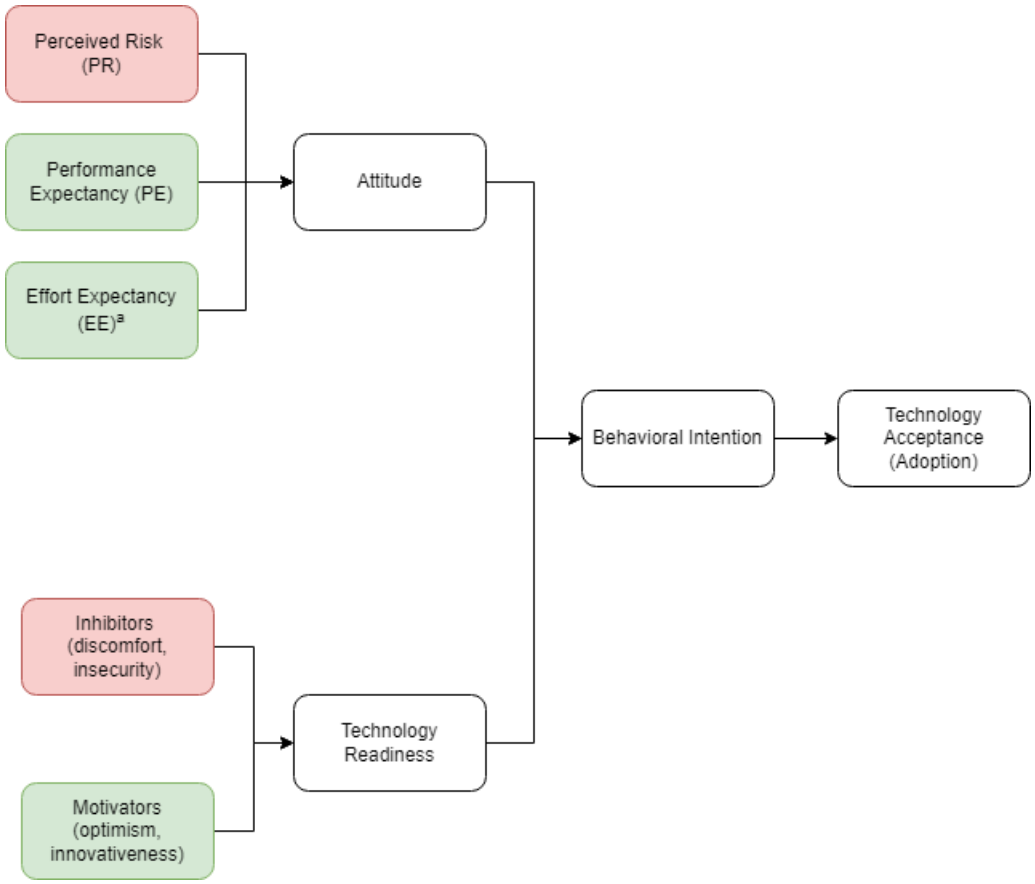

Figure 3. Technology Adoption Conceptual Framework. As Adapted From the UTAUT Model by Venkatesh et al. (2003) and the Technology Readiness Index by Parasuraman (2000).

[a] With Effort Expectancy only taken into account for teachers

The model consists of three variables of the UTAUT model: Perceived Risk, Performance Expectancy, and Effort Expectancy, and two variables of the Technology Readiness Index: Inhibitors and Motivators. It is important to note that the construct of Effort Expectancy will only be incorporated for the teachers since they are the only party that will be interacting with AI programs, whereas the students are only on the receiving end.

Both the UTAUT and Technology Readiness models will influence the user's behavioral intention to choose whether to adopt a certain technology or not. Whereas the UTAUT model visualizes the user's attitude toward a certain technology in context and the Technology Readiness model visualizes the user's technological preferences towards technology in general.

1.3 Research Question

The aim of this thesis is to investigate how students and teachers at the University of Twente react to situations in which AI is used for assessment in an educational setting. To fulfill this, an explorative qualitative research design will be used to gather data. By using scenarios and interviews, information will be gathered on how different situations change participants' respective

technology acceptance. The scenarios require the participant to judge the used AI within different educational contexts. These insights aim to construct which factors affect the adoption of AI assessment adoption. The following research question will be researched in this thesis:

- What is the relation between personal attitudes and technology preferences on the acceptance of assessment AIs in formal Higher Education?

2 Method

This study focuses on how certain scenarios influence someone's behavioral intention (or acceptance) to use AI in an educational context. The study is qualitative research with an exploratory nature. Participants were asked to complete a survey and additional interviews were held with a selected sample of individuals. Ethical approval for this study was given by the Ethics Committee BMS at the University of Twente. The dependent variable in this design was the behavioral intention to use artificial intelligence and the independent variables were performance expectancy (PE), effort expectancy (EE), perceived risk (PR), motivators, and inhibitors.

2.1 Respondents

Participants were gathered through a personal network and convenience sampling within the population of students and teachers currently studying and working at the University of Twente. The study involved three different groups: bachelor's students, master's students, and teachers. In total 20 participants responded to the survey (5 bachelor's students, 7 master's students, and 8 teachers).

For the next round of qualitative data gathering, 12 participants were manually selected to participate in an interview (4 bachelor's students, 3 master's students, and 5 teachers).

2.2 Instrumentation

A survey with both closed and open-type questions was used to gather responses. The survey was divided into three parts.

The first part aimed to collect the participants' informed consent and demographics. The data gathered in this survey was anonymized. The participants were asked to provide the researcher with their educational level. See Appendix A for the full version of the informed consent form.

The second part determined the participant's baseline level of technology acceptance, as adapted from questionnaires used by Damerji and Salimi (2021) and Andrews et al. (2021). It included 15 closed questions (scale of 1-100) on the factors of Performance Expectancy, Effort Expectancy, Perceived Risk, Motivators, and Inhibitors. Appendix B discloses the full version of the survey.

The third part included scenarios that described a plausible situation in which AI was used for student assessment. Scenario planning is a useful technique to prepare for uncertain futures (Gutierrez, Perez, & Munguia, 2022; Amer, Daim, & Jetter, 2013). The scenarios used in the survey were an outline of certain aspects of AI used in an educational context. Their goal was to highlight the implication of certain aspects of AI in an as realistic as possible context. Five factors were chosen to base the scenarios on: efficiency, bias, effectiveness, transparency, and ease of use. Students did not receive scenarios with the ease of use factor since they do not directly use the assessment AI software. Similarly, teachers did not receive scenarios with the bias factor since they are not the direct subject of bias in assessment. Tables 1 and 2 provide an overview of the used factors and scenarios for students, whereas Tables 3 and 4 for teachers. Appendix C discloses the full scenarios. The participants were asked to read the scenarios and give their reactions to a number of open questions.

Table 1

Summary of Used Scenarios for Students.

Scenario	Efficiency	Bias	Effectiveness	Transparency
Language bias	+	-		
Limited transparency		+		-
Strict guidelines	-		+	
General feedback			+*	
Transparency error	+		-	+
Double work	-			+

Note. + indicates a positive effect on a given factor, - indicates a negative effect on a given factor. *Note**. Scenario *'General feedback'* presents a situation in which the AI assessment is correct and effective, but the feedback is not as personal as teacher feedback.

Table 2

Student Scenario Summaries.

Scenario Title	Summary
Language bias	The AI system provides the student with a grade right after the deadline passes. However, the student failed the assignment. Upon further inspection, the student finds out that the system has a bias towards UK English.
Limited transparency	As opposed to possible teacher bias (stress causing teachers to grade stricter), the AI system judges all students equally. However, it does not provide students with a rubric.
Strict guidelines	In order to receive a predicted grade from the AI system, the student has to re-write their assignment in a specific template, otherwise, the system cannot fulfill its function. It will cost extra time for the student.
General feedback	All students receive feedback prompts on their assignments, but the feedback prompts are not as personal as possible teacher feedback.
Transparency error	The AI system provides the teacher with suggestive grades and feedback prompts. Not all feedback prompts are based on correct conclusions. The student notices that the teacher has accepted all feedback prompts and based their grades on the suggestive grade.
Double work	The AI system goes through student assignments and provides the teacher and students with feedback prompts on its decisions. The teacher, however, decides to go through all prompts and takes the usual two weeks to grade, making all the decisions themselves.

Table 3

Summary of Used Scenarios for Teachers.

Scenario	Efficiency	Ease of Use	Effectiveness	Transparency
Rubric incompatibility	-			+
Invisible rubric		+	+	-
Non-efficient interface	+	-		
Impersonal feedback	+		-	

Note. + indicates a positive effect on a given factor, - indicates a negative effect on a given factor.

Table 4

Teacher Scenario Summaries.

Scenario Title	Summary
Rubric incompatibility	The AI system bases a predicted grade on the provided rubric. However, it is not compatible with the rubric and assignment created by the teacher. In order for it to work, the teacher has to re-write these materials.
Invisible rubric	Easy-to-use software has been implemented in Canvas and quickly provides grades for student assignment. However, it lacks transparency on how grades were decided. When the teacher grades an assignment themselves, the two grades show similarity.
Non-efficient interface	Teachers need to invest some amount of time before they can use the AI system. However, the system does bring promises of efficiency when used proficiently.
Impersonal feedback	The AI system can provide student assignments with standard feedback prompts. However, these prompts are less personal than the feedback that the teacher usually gives.

Semi-structured interviews were held with a selected sample of participants. Questions were derived from the UTAUT and TR models, as well as reactions to the responses of the participant given in the survey. The interview guide can be found in Appendix D.

2.3 Procedure

Qualitative data were collected in two phases. In the first phase, participants were asked to fill in the survey. Survey data was collected through Qualtrics and could be answered in Dutch or English. Completing the survey took 15 to 30 minutes. The survey consisted of four parts.

1. Informed consent: Participants were presented with information on how results are collected and processed. In order to continue, participants must give informed consent that their data can be used for the purpose of this study.

2. Demographics: Demographic information on the participants was collected through three questions.

3. Baseline technology acceptance: Participants answered a set of questions that determined their baseline level of technology acceptance. These questions aimed to individually measure the constructs of PE, EE, PR, Inhibitors, and Motivators of AI. (Appendix B)

4. AI assessment scenarios: Participants were presented with a series of scenarios in which AI was used in an educational context. The scenarios presented different contexts and situations, some portraying AI in a negative and some in a positive light. The scenarios differed in certain aspects. After reading the scenario, participants were asked to answer a set of questions that determined whether the scenario influenced their view on AI in education. Bachelor's and Master's students had to complete six scenarios, whereas teachers had to complete four scenarios. (Appendix C)

In the second phase, after data on their reactions was analyzed, a selected sample of participants was invited for semi-structured interviews. In total, 7 students and 5 teachers were interviewed. All interviews were analyzed in Atlas.ti. Due to technological reasons, analysis of one interview could not be done in Atlas, so the analysis was done based on notes taken during the interview.

2.4 Data Analysis

Both qualitative data and quantitative data were generated by the survey instrument. Descriptive statistics were used to calculate values for the technology acceptance levels using SPSS software. The values for Perceived Risk and Inhibitors were reversed. Survey responses were analyzed by in-vivo coding and categorized based on the level of participant acceptance of the scenarios. The codes of the survey responses were used to design participant-specific interview guides.

Based on the theoretical framework, a coding scheme for the interviews was created. This coding scheme existed of five categories: PR, PE, EE, inhibitors, and motivators. Codes were assigned to responses in the interviews that depicted the variables and later compiled into different themes under that same variable. Themes were derived whenever at least three participants mentioned a similar occurrence. In total, 13 themes were deducted from the student interviews and 10 themes were deducted from the teacher interviews. A single researcher performed the coding, therefore no inter-coder reliability was calculated. The full coding scheme can be found in Appendix E.

3 Results

In the following section, the results of the student participants will be discussed. This section is divided into survey responses and interview analysis.

3.1 Student Survey Responses

In total, 12 students filled in the survey (5 Bachelor, 7 Master). Tables 5 and 6 give an overview of the bachelor's and master's students' scores on the baseline questionnaire, respectively. Bachelor students are identified as 'B(n)', and Master students are identified as 'M(n)'.

Table 5

Bachelor Students Baseline Scores on a 1-100 Scale (n = 12).

	Bachelor Students					Total	
	B1	B2	B3	B4	B5	M	(SD)
Perceived Risk	42.0	28.8	55.0	38.3	41.7	41.1	(9.4)
Performance Expectancy	74.3	61.0	61.0	60.7	73.3	66.1	(7.1)
Inhibitors	80.5	70.0	61.3	42.0	33.8	57.5	(19.4)
Motivators	41.3	41.3	44.3	38.8	67.5	46.6	(11.8)
Average	59.7	51.0	55.0	44.3	53.6	52.7	(5.7)

Bachelor students scored highest on Performance Expectancy and lowest on Perceived Risk. On average, B1 (59.7) scored highest on the baseline questionnaire among Bachelor students, followed by B5 (55.0). The lowest scoring Bachelor student was B4 (44.3).

Table 6

Master Students Baseline Scores on a 1-100 Scale (n = 7).

	Master Students							Total	
	M1	M2	M3	M4	M5	M6	M7	M	(SD)
Perceived Risk	40.0	49.3	18.3	44.3	44.3	50.7	60.0	43.9	(12.9)
Performance Expectancy	73.3	64.0	81.7	29.3	64.0	68.7	65.0	63.7	(16.4)
Inhibitors	52.5	74.3	39.0	54.0	48.5	50.0	50.0	52.6	(10.7)
Motivators	40.0	45.3	56.3	57.0	88.8	51.0	46.3	54.9	(16.1)
Average	50.7	58.4	48.6	47.5	62.4	54.4	54.2	53.8	(5.4)

Master students scored highest on Performance Expectancy and lowest on Perceived Risk, like the Bachelor students. For the Master students, M5 (62.4) achieved the highest average score, followed by M2 (58.4). The lowest scores were dedicated to M4 (47.5) and M3 (48.6). All four participants accepted two out of six scenarios. M3 and M5 even accepted the same set of scenarios.

Bachelor students scored lower on the Motivators factors.

Table 7 gives an overview of the acceptance of the provided scenarios.

Table 7

Summary of Scenarios as Answered by Students (n = 12).

Scenario	B1	B2	B3	B4	B5	M1	M2	M3	M4	M5	M6	M7
Language bias												
Limited transparency												
Strict guidelines												
General feedback												
Transparency error												
Double work												

Note. Cell colors indicate acceptance of a scenario as follows: green = acceptable, red = unacceptable.

It is interesting to point out that while students scored high on the baseline questionnaire this was not an indicator of acceptance of scenarios. The two highest-scoring Bachelor students B1 and B5 accepted four and one scenarios, respectively. On the other hand, the lowest-scoring Bachelor student, B4, accepted two scenarios. Similarly, among the Master students, the highest-scoring students M5 and M2 both accepted two scenarios. Most scenarios were accepted by M7, who had an average score.

Most Accepted Scenarios Two scenarios received an equal response. One of which was the scenario was 'General feedback'. Nine out of twelve participants accepted the scenario. Participants valued feedback and the setback of standardized feedback was not negative enough for them to deem the scenario unacceptable. However, participant M5 stated: *'This would be acceptable as long as a clear justification/reasoning is provided by the AI to explain the right answer and availability of more information making us aware how to reach the correct resolution.'*. The three participants that did not accept the scenario were mostly troubled by the lack of their teacher checking their assignments and the overall lack of personal feedback.

The second scenario was 'Double work'. The leading cause for participants to accept the scenario was the fact that their teacher made the final decision on the grades. The three participants that did not accept the scenario commented that the introduction of the AI service did not increase efficiency at all: *'The few weeks the teacher had to spend might've been a lot better for them than the time they would normally take to grade the assignments. However, even if this is the case, an AI that needs to be fully babysat by [a teacher] does not help anyone. (M3)'* However, some participants saw it as a good chance to test the AI system. Scenario 6 was more likely to be accepted by Bachelor students.

Least Accepted Scenarios The least accepted scenario was 'Language bias'. This scenario dealt with bias in the AI system. Participants voiced their worries that the guidelines were too strict and that this would never happen when a teacher would assess an assignment. Especially since the AI software assessed based on language mechanics instead of content. Responses to how this system could be used correctly were mixed, ranging from communicating these stern guidelines and thus avoiding situations like this to implementing a system in which students can let their teacher go over their grades.

The second least accepted scenario was 'Transparency error'. Only participant B1 accepted the scenario, stating: *'If it is sufficient, I don't really care'*. The other participants, however, did not accept a scenario in which an AI system would cost them their grades by making mistakes. One participant commented: *'No, because the software should only help the professor [do] his/her job and not take his/her place. In this scenario, it seems that the professor didn't even look at what the software did. (M6)'*

The scenario 'Limited transparency' received negative responses from 9 out of 12 respondents. All 9 participants said that they needed insight into grading decisions before they would be willing to accept the grade. Participants either wanted in-depth feedback so that they could learn from their mistakes or insisted upon receiving a rubric to see how the AI system calculated their grades. Of the 3 participants that accepted the scenario, participant B1 was not disturbed by the lack of transparency, while the other two would like some level of transparency whenever they would not agree with the provided grade.

Mixed Responses Interestingly, the reactions toward scenario 'Strict guidelines' were mixed. Participants were willing to accept the scenario if the guidelines that the AI system used to grade were shared beforehand, however: *'If it just says that it is not satisfactory then rewriting will not be very helpful because after all, I am the same person who might write the same thing (M7)'*. Participants that were not keen on the scenario, commented on the inflexible guidelines and that variation was a natural process when writing assignments. *'Papers don't usually follow the exact guidelines. If the software only can follow these strict guidelines, I don't think it should be used. (B4)'*. Scenario 3 was more likely to be accepted by Master students.

3.2 Student Interviews

To further investigate the findings of the survey, the interviews aimed to gain knowledge on why participants made certain decisions. This qualitative analysis was based on the five factors of the theoretical framework: Perceived Risk, Performance Expectancy, Effort Expectancy, Inhibitors, and Motivators. These will be discussed in the order of the framework.

Participants B1, B2, B4, B5, M1, M3, and M4 were interviewed for this section.

3.2.1 Perceived Risk

Concerning this category, four themes were recognized to be associated with an increased sense of perceived risk:

- The inaccuracy of assessment
- No room for discussion
- Inflexible grading guidelines
- The decrease in the quality of education.

Inaccuracy of assessment The most common theme was the inaccuracy of assessment, as mentioned by six out of the seven interviewees. A common worry was the occurrence of bias in the form of false negatives and false positives in grading, i.e., that the AI will interpret things wrong and provide a wrong grade as output. Interviewees were especially worried about the system making mistakes in the sense that it would provide a more negative grade than a teacher would. In most cases, interviewees expressed a negative view of the AI system whenever it would not be equal to the assessment they receive from their teachers, as expressed in the following quote:

> Q1: "I think that what I really find important is that the AI [assessment] is equal to the professor, so that I won't be disadvantaged by being assessed by the AI." [B4]

As far as correctness of assessment is concerned, students express that they think that the AI system would not be able to handle text that falls outside of its training dataset and therefore will not be able to accurately match things that would be considered original or a variation on the correct answer by a human. Another way in which this bias would affect students was through not recognizing dyslectic or technologically disabled people. In the following quote, this issue is addressed as 'originality':

> Q2: "I don't know if there is a specific name for that kind of bias, but that variation isn't possible. So, answer A is right and if you fill in answer B, which would also be considered correct, will be marked as wrong. A teacher would pick up on that in a way that it wouldn't be included in the answer key, but technically, it is correct. An AI system won't be able to do that with an answer it has never seen before." [M3]

No room for discussion The second most common the perceived risk was the inability to discuss grading decisions with an AI program. Interviewees could not imagine that it would be possible to enter a discussion with the algorithm, especially with the current developments of AI. As was found in the interviews, interviewees want to have the option to discuss disagreements within the assessment form and do not think AI services are capable of that.

> Q3: "If I look at the current state of AI. [..] For example, chatbots are difficult to have a discussion with and I personally think it is important to be able to discuss something when you don't agree with the AI's decisions." [B2]

Interestingly, the participants that brought these arguments forward all accepted the scenario in which the AI service provided them with basic feedback prompts. So, as long as the AI system is correct or treats students to their advantage, they are willing to accept its decisions.

As a response, interviewees expressed the wish that either a person always looks over the grading decisions that the AI system has taken or that there is an option for a second opinion from a teacher when students do not agree with their received grade.

> Q4: "I would like it that if I have some of my answers wrong that there is some point of contact where I can get an explanation [on certain grading decisions]." [B5]

However, these worries were not expressed on assignments that include calculations. Only written papers and reports were mentioned as assignments where students would want to discuss grading.

Inflexible grading guidelines The third most common theme was the inflexible grading guidelines the algorithm would uphold. All seven interviewees mentioned that they needed clarity and transparency when an AI would make grading decisions. This clarity would come in the form of feedback and assessment models (rubrics). If an AI program were to grade a student, it would need to explain why an assignment has come to a certain grade. If a grade is not accompanied by feedback or a rubric, interviewees found that the grade would be unacceptable. The following quote highlights this need for rubrics and an explanation of certain grading decisions:

> Q5: “Everyone should be able to understand why certain decisions were made: how it was graded, why certain things were marked as mistakes, etc. Otherwise, people will ask themselves why something is marked as wrong." [B1]

The interviewees that expressed concerns about guidelines, stressed that an AI program would probably be more inflexible than a teacher would. It was often mentioned that a teacher and an AI program could have different perspectives on what is important in an assignment and therefore, come to a different grade. In most cases, interviewees thought that the AI would uphold different, and stricter, guidelines than a teacher would. As a result, an interviewee mentioned that students might adhere more strictly to guidelines whenever graded by an AI program.

> Q6: “If I know that I will be assessed by an AI program, then I will keep strictly to the provided guidelines and probably write with less freedom." [B5]

However, it should be pointed out that M3 and B5 differed in their reactions to the scenario in which the AI system worked with stricter guidelines. While B5 dismissed the scenario based on the way they might get points deducted for not adhering to the assignment structure. M3 argues that the guidelines were provided to the students beforehand and, as long as they were clear, the result would be acceptable. So, there is a difference in the level to which a student must adhere to guidelines and whether students are willing to sacrifice that level of freedom in their writing.

Decrease of the quality of education The fourth theme was the fear that AI assessment could decrease the quality of education. Five of seven interviewees mentioned their worries that introducing AI assessment might risk the loss of quality in education. Three interviewees mentioned that they would think the value of their diploma would be decreased if AI assessment were to take over grading.

> Q7: “It sounds rather extreme when you think that you finished your studies and didn’t receive your grades from a person, but from an AI service." [B1]

However, interviewees often mentioned concerns relevant to their studies and values. For example, the misfit of AI assessment in qualitative-oriented assessment or the missing empowerment of students. Multiple concerns were raised on the field of AI assessment and how it would affect student education.

3.2.2 Performance Expectancy

Within this category, four themes were recognized as a possible added value to education:

- Receiving feedback,
- Faster grading,
- Decrease of human bias,
- Time efficiency for teachers.

Receiving feedback Receiving feedback was mentioned most often by all seven participants as an effect of using AI for assessment. Three out of seven interviewees mentioned that AI might be able to provide more extensive or detailed feedback that teachers might not give. It was seen as added value if AI could always generate feedback. One interviewee even mentioned that if the provided AI feedback was detailed enough, they would have no need for a second opinion from their teacher.

However, four interviewees mentioned that they would always appreciate teacher-given feedback more. Especially with open-ended questions and essay-type papers, four out of seven interviewees did not see added value in AI-generated feedback. The following quotes illustrate why:

> Q8: "I would not like it if it [AI-generated feedback] would be used for papers because it is free input what you put in a paper. So I think the feedback would not overrule that of a teacher. [..] " [B5]

Faster grading The second theme found for performance expectancy was that of faster grading, as mentioned by five out of seven interviewees.

> Q9: "Added value would be to know if you are going in the right direction before you hand something in for a grade. [..] That helps with finishing the last part. Like 'Am I on the right track?' and to have more insight into where your grade is heading." [B2]

The cases that did feature the importance of faster grading were the possibility to know whether you had to take a resit earlier, knowing where your assignment's grade was heading before the deadline, and that it might be more time efficient for teachers.

However, it should be noted that three out of seven interviewees did not see the direct added value for an increase in grading speed in their current situation. It did not have any effect on their workstyle and they did not experience a need for faster grades.

Decrease of human bias The third theme is the decrease of human bias. While bias is a heavily discussed aspect of AI, six out of seven interviewees mentioned that the objective evaluation of AI could remove possible teacher bias from educational assessment. This is because interviewees thought that the AI program would treat every student with the same grading system.

There were mentions of human biases, e.g., that teachers would suffer from interaction with their emotional state when grading. Four out of the seven interviewees mentioned something about

subjectivity in grading by their teachers. While the impact of this bias is not large, it is noticeable enough for students.

> Q10: “It is different than being graded by a teacher. Of course, it isn’t 100 percent foolproof, everything is based on an opinion or the mood of a teacher." [B1]

Time efficiency of teachers The final theme was the time efficiency of teachers. Six out of seven interviewees mentioned that the quality of education could be increased when teachers would be (partially) alleviated of their assessment duties. They said that teachers could invest more time in the preparation of their lectures, provide more in-depth feedback, and focus on content rather than the structure of assignments. However, it is interesting to note the difference in takes on this subject in these two quotes:

> Q11: “It [faster grading] does not do much for me, but I reckon that the teacher will have more time for other things, like for their students." [M3]

> Q12: “I would like to see that it [AI assessment] would enable more people to learn. For example, that a professor can teach a class double its normal size or that classes become cheaper so that more people can access education." [M1]

An interesting phenomenon is observed in the scenarios, where most participants did not accept the scenarios that actually increased the efficiency of teachers (scenarios 1 and 5), but they did accept scenarios in which the teacher would have to spend equal or more time on assessment. So, while they voice better time management for teachers, they are not choosing the scenarios in which time management is actually improved.

Nevertheless, students did not think that this would have a direct effect on them. So, their optimism on this theme was mild.

3.2.3 Inhibitors

For this category, three themes were found:

- Lack of trust,
- Lack of human interaction,
- Dependence.

Lack of trust The most common theme was the lack of trust in AI services. Five out of seven interviewees called upon a distrust towards AI. Multiple reasons for this distrust were found, the first being that the situation of AI independently assessing student work was a new concept to interviewees.

The trust in AI is feeble. One interviewee mentioned that they had a bad experience with an AI service assessing them before:

> Q13: “We worked with [software program] before for a mathematics course. I felt okay with the software grading me, but in the end, other students found three errors in the way the software assessed us. Three mistakes happened to a lot of people. So, if AI can be without any errors, then I don’t see a problem. But until then, I would like it if there is a teacher to look over the output. Because if we [the students] have to check for mistakes, I would not like that." [B5]

So, as long as AI services can make mistakes, interviewees mentioned that they would rather have a human involved in the grading and assessment process. In the interviews, it came to light that students have an easier time putting trust in someone with experience, e.g. they have a diploma that shows their mastery of knowledge, than a software program.

> Q14: “If an AI program would give me a 6, I would immediately start thinking: ’Okay, you can say that, but what process did you go through to get to that grade?’. While, if a teacher would give me the same grade, I would be more likely to think that they are right. [..] I would look more critically to what an AI does, so I think it’s somewhat dangerous to have AI grade people." [M3]

Lack of human interaction The second theme was the lack of human interaction in AI. Four out of seven interviewees mentioned that they would miss the human interaction they have with their teachers if an AI program would start grading their assignments. Something that came up multiple times was that interviewees stated that they followed a course for the teacher and would then like to be graded by that teacher, not an AI program:

> Q15: “Because it feels like you have expectations when you come into class. If the teacher is teaching the entire class, you expect to receive something that suits his style or her style. But if eventually all your hard work is graded, for instance via AI then it feels a bit like betrayal." [M4]

Interestingly, scenario 4 dealt with the replacement of human feedback with AI-generated feedback. Most participants accepted the scenario, stating that they favor quick feedback over human-given feedback. This shows a discrepancy between interview quotes and survey responses.

Dependence The third theme found was dependence. Five interviewees mentioned that they would not like to be dependent on AI assessment in education. As mentioned before, these interviewees like to have a human involved in the AI-generated assessment for the sake of checking the output. However, this does not mean that these five interviewees are against the use of AI in assessment:

> Q16: “If it can make the tasks of teachers easier, that would be really nice. Maybe teachers do not have to check grammar or those kinds of things, but they can focus on the main points of an assignment, and the AI can do the rest. Maybe a quick check through the things AI was doubting about." [M3]

The dislike of dependence on AI assessment mostly stemmed from the fear that AI-generated decisions would be accepted without thinking. Three out of seven interviewees mentioned that they need a human to look over AI decisions before accepting the grade or assessment. However, one mentioned that it would be okay to be dependent when using AI would provide them with positive consequences.

3.2.4 Motivators

Within the category of motivators, two themes were found:

- Trust
- Technology potential.

Trust Trust was mentioned by four interviewees. A leading factor in trust was mentioned to be prolonged exposure to positive interaction with AI programs. If grades provided by AI software are accurate and no mistakes are made, then students are more willing to be more trusting.

> Q17: "Maybe if AI programs are further developed, it [worrying about whether the output is correct] might not even matter anymore. Because people might have enough trust in AI and think 'Oh, last time the teacher agreed with the output and I did as well' and after a while, they simply start to accept AI." [B4]

It is interesting to see that these two participants mention the factor of trust and how it can be increased, since most participants did not accept more than three scenarios. This calls for more positive experiences with AI technology in assessment.

However, trust is the opposite effect as distrust in Inhibitors. People are not used to AI services determining grades and assessments, but one of the ways to grow trust is prolonged and positive exposure to its effects. One interviewee mentioned that trust in AI could be increased by not only implementing it in one university but to involve multiple universities in its effort. In another light, one interviewee mentioned that they would trust the university enough to implement a working AI service.

Technology potential The second theme was the technology potential of AI services. Four interviewees mentioned elements of AI that they would like to see implemented in education. For example, one interviewee mentioned that they saw the benefits of implementing AI assessment in scientific studies.

> Q18: "I think that for hard sciences, like mathematics or a language, where your answer is either right or wrong [..] there is some implementation possibility for AI." [M1]

Interviewees agreed that AI assessment is suited for multiple choice, fill-in-the-blanks, and other question types where it is easy to program the 'right' answer. There seems to be some confusion about when AI assessment takes place since grading closed-type questions is not a task done by AI systems.

However, there are some limits to students' motivation to use. Two interviewees mentioned that they would like to see students' consent before AI assessment would be implemented in their classes and that the AI program should be properly tested before it would be used to determine student grades.

> Q19: "I think the distrust is there because it [AI assessment] is rather new to everyone. [..] For me, it is important that everything is tested thoroughly before it would be implemented. [..] Especially in the first phase, it is important to test the program." [B4]

3.2.5 Summary Students

Table 8 depicts a summary of the findings during qualitative analysis for the students.

Table 8

Summary of Student Findings.

Perceived Risk	Performance Expectancy	Inhibitors	Motivators
Inaccuracy of assessment	Receiving feedback	Lack of trust	Trust
No room for discussion	Faster grading	Lack of human interaction	Technology potential
Inflexible grading guidelines	Decrease of human bias	Dependence	
Decrease of quality of education	Time efficiency		

3.3 Teacher Survey Responses

In total, 8 teachers filled out the survey. Table 9 gives an overview of the teachers' scores on the baseline questionnaire.

Table 9

Teacher Baseline Scores on a 1-100 Scale (n = 8).

	Teachers								Total	
	T1	T2	T3	T4	T5	T6	T7	T8	M	(SD)
Perceived Risk	53.0	50.0	50.0	71.0	33.3	58.3	51.7	68.3	54.5	(11.8)
Performance Expectancy	47.0	91.7	66.7	50.7	58.3	75.3	65.0	68.3	65.4	(14.2)
Effort Expectancy	56.7	100	70.0	56.7	66.7	67.3	66.7	43.3	65.9	(16.3)
Inhibitors	54.0	40.0	57.5	72.8	18.8	64.3	48.8	22.5	47.3	(19.2)
Motivators	51.5	93.8	65.0	64.5	75.0	42.5	61.5	48.8	59.7	(11.0)
Average	50.8	79.1	61.8	63.1	50.4	61.6	58.7	50.3	59.5	(9.6)

Teachers scored highest on Effort Expectancy, followed closely by Performance Expectancy, similar to the students. Teachers scored lowest on the Inhibitors factor. On average, T2 (79.1) scored highest, followed by T4 (63.1). The lowest-scoring teacher was T8 (50.3), followed closely by T5 (50.4) and T1 (50.8). While teachers have a higher

Table 10 gives an overview of the acceptance of the provided scenarios.

Table 10

Summary of Scenarios as Answered by Teachers (n = 8).

Scenario	T1	T2	T3	T4	T5	T6	T7	T8
Rubric incompatibility								
Invisible rubric								
Non-efficient interface								
Impersonal feedback								

Note. Cell colors indicate acceptance of a scenario as follows: green = acceptable, red = unacceptable.

Interestingly, T2 was the teacher with the most accepted scenario prompts, as opposed to T1 with two accepted scenarios. As the teacher with the lowest score, T8 was surprisingly not the teacher with the lowest number of accepted scenarios, which was T6.

Most Accepted Scenarios Six teachers accepted the scenario 'Impersonal feedback'. As long as the feedback was theoretically appropriate, these six teachers thought it would be a sufficient implementation. An important note is that three teachers mentioned that they would edit or add personal feedback to the provided prompts as a way to stimulate student learning. T5 highlighted: *'The AI should tell me about the cases in which my intervention might be needed because of high uncertainty.'*. However, T8 countered with: *'A more generic feedback prompt can still be very valuable, as long as students know how to deal with them.'* The two teachers that did not accept the scenario placed more value on personal feedback, *'A trade-off between efficiency and thoroughness may not be good. (T3)'*.

Least Accepted Scenarios One of the main reasons for not accepting the scenario 'Rubric incompatibility' for the teachers was the amount of extra work it would take to re-write their assignments into a state the AI program could use. *'This is not acceptable the AI should adapt to me not the other way around'*, was the statement of T5. If teachers did not get enough time to get familiar with the software, they were less likely to want to use it: *'Lacks efficiency as it does not help you save time. (T8)'*. The two teachers that did accept the scenario mentioned that if the program could benefit them in the future, they would be willing to put in a little extra work.

Mixed Responses Among the mixed responses was the scenario 'Invisible rubric'. All respondents valued transparency in the grading process, but five teachers found the accuracy of the provided predicted grade enough to overthrow the lack of transparency. However, it is important to note that the teachers that accepted the scenario still stated that they would go over the assignments themselves so that they are able to elaborate on the grading. One of the teachers compared an accurate AI system to a consistent student assistant, meaning that they see the AI program in an

assisting role. The teachers that did not accept the scenario stated: *'You can never use a black-box tool for grading, it is essential to know how it operates before relying on its results.' (T6), 'I would like the AI to also explain its grade decision based on the rubric, without me having to sample its results and hoping they overlap. (T2)'*.

In the scenario 'Non-efficient interface', again the reason for not accepting was the amount of extra work it would take to get to know the AI program. However, it should be noted that the teachers that did not accept the scenario were also keen to point out that they would need an intuitive and easy interface. T5 commented: *'Usability and timing are essential I can not expose students to a new approach that I do not know.'*, which aligns with the teachers' need for transparency. The three teachers that were accepting of the scenario mentioned that a learning curve is expected when implementing new software and: *'If the plan is to use it for multiple years and for multiple courses/modules, then it is worth the time investment. (T2)'*.

3.4 Teacher Interviews

Participants T1, T2, T3, T4, and T8 were interviewed for this section.

3.4.1 Perceived Risk

Within this category, four themes were recognized:

- The quality of education,
- Risk of bias,
- Transparency issues,
- Replacement of staff.

Quality of education All five interviewees mentioned something about how they thought AI would affect the quality of education. While highly hypothetical, the opinions of the interviewees ranged from neutral to negative views on the implementation of AI in assessment:

> Q20: "It is really difficult to say and conclude that it will positively influence education. It might give me more time to take care of other courses, but I don't think that those courses will be better or worse with the implementation of AI." [T1]

> Q21: "So imagine that we are using this kind of system. And as soon as we also think that assessments are not made by the professors, the teachers, or the lectures then this can also change the way people think, and it may not always be in a good way. [..] So you can think that now, for example, teaching or assessment is changing with every teacher in a smaller minor or major way. But then when you start using those systems, you may see that it won't change anymore, so this can also can be a barrier for better teaching skills, better teaching methods, and so on." [T4]

The biggest consequence of assessment is the grades that students receive. Thus, interviewees were less keen to perform summative assessments when using an AI service. A more positive note was attached to using programs for formative assessment.

Teachers mention that assessment is an important point of input for their assignments and lectures. Introducing AI assessment could therefore influence these input:

> Q22: “The input you get from assessment is something you can use for the next set of lectures; explain something again. It gives you insights into what was or wasn’t clear for students." [T8]

Assessment provides teachers with valuable feedback and if assessment gets fully automated by using AI services, teachers might start missing out on important insights.

Risk of bias The second theme was the risk of bias occurring within AI assessment, which was mentioned by three out of five interviewees. In the interviews, it comes forward that teachers view their way of grading as ’the right way’. This means that any assessment scheme differing from theirs will be faced with criticism.

All three interviewees were involved with the grading of their students and mentioned some problems that would occur when AI assessment would be implemented. If an AI program were to grade certain assignments, it would not be capable of reading ’in between the lines’ and might punish students for answers that a teacher would reward them points for.

> Q23: “I’m afraid that it [assessment] will become rather objective. That the system will say that something is not included in writing, while the students actually described it sufficiently and show that they understand the material, but did not use the right words." [T8]

Quote 23 highlights an important issue, where teachers focus on the progress a student makes in the run of a course. It was mentioned by two interviewees that they might grant students extra points when they recognize their own lectures in the student answer.

> Q24: “It might be ’my fault’ that the student didn’t get the question. So yeah, I think I might be less strict in that case and that it could be a bias. [..] I would still mark it as wrong, but, in case the question is worth 8 points, I would maybe decrease it with 2 points instead of 3." [T1]

Transparency issues The third theme is the transparency of the AI service. All five interviewees mentioned transparency in their interviews and thought that it was a compulsory aspect of AI when it would be implemented in assessment.

> Q25: “The important thing is to be able to understand why there’s this result at the end, and then, for example, a rubric is a good example for that. So that’s why the end grade is collected right? So also it might be some textual feedback. I don’t know how it will work, but. If there was OK, this grade is given because of this problem, so it can be a standard message of course, but at least you have an idea why." [T4]

While transparency seems to be an important theme for teachers in assessment, five out of eight teachers accepted scenario 2, where transparency was reduced in order to increase efficiency. Where teachers state in the interviews that they value insights into the AI grading process, they were still accepting of a more lacking AI system.

Without transparency, interviewees would not be able to accept the usage of AI programs in assessment. Their trust is built upon transparency. However, the issue of transparency can be covered by using rubrics within the AI program. If the output of the AI program would be a rubric that justifies all grading decisions, teachers would be more willing to accept the outcomes.

Replacement of staff The fourth theme found was the replacement of staff. Three interviewees mentioned this theme and all agreed that replacing human staff with technology would negatively affect assessment, and therefore education. While AI services can aid people, they should not completely take over their tasks.

> Q26: "I have designed a digital learning environment and many reactions were 'Do you want to replace me?', and my answer was always: 'No'. The only way in which this will work is if there is a teacher around to help out. Moments in which exceptions occur. An AI can cover a lot of substances, but not every student is the 'average' student. [..] So it's nice to have this kind of cooperation." [T8]

This calls for hand-in-hand cooperation between people and AI technology.

3.4.2 Performance Expectancy

Within this category, three themes were recognized:

- Positive time management
- AI in a supporting role
- AI for specific assessment types.

Positive time management All five interviewees mentioned that the implementation of AI software for assessment would improve their time management. Assessment was seen as a 'dull' part of the job and was mostly located lower on teachers' task lists. Teachers mentioned that especially in assessment, they would be willing to hand over some of the work.

> Q27: "I think it can really save some time. And as I said [..] for most people it's kind of a dull job to make assessments. So I can use that time for more productive things." [T4]

However, the time that would be freed would not always go back into the course:

> Q28: "I notice that I already take enough time to prepare my lectures. So, one week in advance I will start working on them. That is already enough time." [T1]

Two of the five interviewees mentioned that they already took enough time for their courses and would not need AI support solely to improve the quality of their courses. The other three interviewees mentioned that part of their freed-up time would be invested into their courses, by restructuring their lectures or engaging in more interaction with their students.

AI in a supporting role All five interviewees mentioned that they would see AI programs in a supporting role when used for assessment. Two teachers specifically mentioned that they would only see an AI service as another version of a teaching assistant. While this shows some level of trust toward the AI service, two teachers mentioned that they checked their teaching assistant's work to make sure everything is in order.

> Q29: "Sometimes people ask me why I do not let my teaching assistants grade, but I think that a teacher should provide their students with a grade. It's complex. And while I think my teaching assistants are capable of helping, there are a few things where I think they might not be capable of grading [..] the more complex grading situations, for example." [T1]

In a similar way, teachers would not feel comfortable putting an AI program in charge of grading students' work. Summative assessment was mentioned to be 'too definitive' and two interviewees would rather use AI services for formative assessment. This does not mean that they do not trust the technology, they simply prefer to be in charge of the grade themselves.

This level of support that the AI service should provide differed per interviewee:

> Q30: "The system can, for example, flag down certain assignments that it thought were difficult to grade." [T8]

> Q31: "If the answer to a question isn't correct, then it becomes too difficult for an AI service. Then it isn't straightforward anymore. In that case, I would like to look at it myself. But if an AI could tell me the fully correct answers, then I would trust that." [T1]

AI for specific assessment types Added value in assessment for certain assessment types was mentioned by all five interviewees. The overall consensus was that AI programs would be skilled enough to deal with question types that have clear right answers. However, the opinions of interviewees started to differ when it came down to assessment types that deal with text interpretation.

> Q32: "I think that when the questions become more complex, so when there is no clear right or wrong or a student makes a mistake in the last sentence that undermines their answer [..] I think that is the point where AI is not ready for." [T1]

> Q33: "And at least I guess it can give an overall idea. [..] I imagine that if it tells me OK, this is the quality that you can expect from this assignment, then it's easier to assess because of one of the things for an assessment." [T4]

When interviewees mentioned the type of assignments they did find suitable for AI assessment, only one thought that present-day AI programs were capable of grading short papers (500 words max). Other mentioned types of assessment were: closed book exams with short-entry answers, recognizing themes and keywords, and objective measures in text.

3.4.3 Effort Expectancy

One theme was identified for the category of Effort Expectancy.

- Inexperienced technology usage.

Inexperienced technology usage Especially with AI programs, three out of five interviewees mentioned that they did not have any experience using this kind of technology in their field of work, especially with assessment. This caused some level of discomfort, as teachers become unsure of how to use an AI program. Problems that could arise are the misuse of technology and increasing insecurity with technology.

> Q34: “My experience is that often help is offered, but that we, as teachers, do not know about that. So we have no clue where to go." [T8]

The hypothetical situation that assessment would be moved to the AI domain was often compared to the move from Blackboard to Canvas. A lot of teachers experienced difficulties but were helped by student assistants that guided them through the program. A point of contact seems to be a likely solution for AI services as well.

> Q35: “I would like to be able to go into conversation with someone. Maybe address some factors I would like to see implemented into the program or discuss my [technology-related] insecurities with." [T1]

One of the interviewees that did not express any discomfort with new technology was T8, who also expressed high scores on the scale of Effort Expectancy.

3.4.4 Inhibitors

No major themes were identified as inhibitors for AI program usage. However, some resistances against AI assessment were mentioned that could be of importance for further implementation.

One interviewee called upon the view of the students, that they would not want to be graded or judged by an AI service. Another point of resistance would come from the lecturers themselves, as two interviewees mentioned that they could see more work coming from the implementation of AI assessment. For example, they would have to adapt their assignments to fit the AI service. One interviewee even mentioned not being willing to go the length to adapt their course structure for the benefit AI assessment.

> Q36: “So of course if I do multiple choice questions it would be much easier for me to assess, right? But then I don’t think it is suitable for my courses to do something like that. Neither exam, nor multiple choice questions are, I think, a really good way to assess the courses that I am giving. So I wouldn’t do a multiple choice [exam] instead of an assignment to fit to an AI." [T4]

3.4.5 Motivators

Two themes were identified as motivators:

- Trust
- Seeing AI as a colleague.

Trust Trust was mentioned by all five interviewees. Similar to students, a leading factor in positive trust was the prolonged exposure to positive interaction with AI programs. However, opinions were divided as to what extent trust would go.

> Q37: “Maybe after a couple of years, when trust reaches 100 percent and grades are good enough, the program might give out grades for small assignments. Not assignments that determine the end grade of a course." [T2]

> Q38: “I would let an AI program grade that [computational questions], but whenever an answer is not right, I would like to look at it myself." [T1]

The differing opinions between T1 and T2 are visualized in their opposite scenario responses. T2 also scored high on the scale of Motivators and thus exhibits a more accepting nature toward technology.

Whenever an AI program would state that something is completely right or wrong, interviewees were willing to already put trust into thinking that would be the right answer. Everything in between differed. One interviewee would not allow an AI program to grade complex questions, while the others were more willing to give it a try. Two interviewees even made a proactive comment that they would want to test whether an AI program would be capable of providing accurate grades.

Seeing AI as a colleague Three out of five interviewees mentioned an interesting concept that they would, in time, view the AI assessment program as a colleague, rather than a tool or student assistant. As expected, these three interviewees did not see AI merely as a supporting tool (see Performance Expectancy).

> Q39: “If it works correctly, it will likely change from student assistant to a colleague." [T8]

> Q40: “Why not? People might say that it is an artificial system. Why should there be such a big difference between a human and a system that is really good at its thing if we come to the same conclusion? There is no problem there." [T2]

Interestingly, T8 shows one of the lowest scores on the scale of Motivators and is still willing to accept AI technology.

3.4.6 Summary Teachers

Table 11 depicts a summary of the findings during qualitative analysis for the teachers.

Table 11

Summary of Teacher Findings.

Perceived Risk	Performance Expectancy	Effort Expectancy	Inhibitors	Motivators
Quality of education	Positive time management	Inexperienced technology usage		Trust
Risk of bias	AI in a supporting role			Seeing AI as a colleague
Transparency issues	AI for specific assessment types			
Replacement of staff				

4 Discussion

The following chapter will highlight the most important findings. First, student results will be discussed, followed by teacher results and resulting in a shared conclusion.

4.1 Students' Attitudes and Technology Preferences on AI Assessment

Participants saw risks in AI assessment, most commonly, in form of inaccuracy of assessment. Especially, when the AI service would give students a lower grade than a teacher would, the students expressed their dislike for the involvement of AI services in assessment. This ties closely to the request that AI assessment should not solely be in control of grading. Instead, all students were in favor of having some point of human contact through which they could go into discussion about the given grading decisions. This agrees with findings by Kumar and Boulanger (2020), in which they stated that AI assessment should always go hand-in-hand with human cooperation to empower students as well as validate the scoring. Giving full responsibility to AI services is not on the table in the foreseeable future.

Students found transparency an important component of assessment and would therefore not accept a system that is not clear in their grading decisions. This agrees with previous literature, in which system explanations increased trust in the student population (Conati, Barral, Putnam, & Rieger, 2021; Ooge, Kato, & Verbert, 2022). For students' empowerment, the system should be explainable, justify its outcomes to the reader, and communicate system errors (Adabi & Berrada, 2018).

It is important to note that student behavior might change according to the assessor. In the interviews, students mentioned that they might change the way they write assignments if they would know that they would be graded by an AI service. A study by G. Zhang, Raina, Cagan, and McComb (2021) showed that the interference of AI assessment had differing effects on low and high-performing teams. Opinions on whether this was acceptable or not differed per interviewee: where some thought it would negatively affect their writing, while others did not mind the strict guidelines as much. However, Brownell (2022) concluded that deception about the assessor will not positively influence adoption rates.

Performance Expectancy (PE) was found to be an area of discussion among interviewees. While PE is a strong predictor of technology adoption, most themes were not as strong when compared to the other factors (Andrews et al., 2021). However, the ability of AI services to provide feedback was a recurring and strong theme. The student participants were divided over the ability to provide feedback by AI services. A leading factor is the level of detail that AI-generated feedback can provide. This is comparable to the results of Chen (2022), in which students mentioned that they preferred AI-generated feedback for grammar and other minor text input feedback and teacher feedback for the overall flow and context of the assignment. Another study by Han and Sari (2022) showed that the combined feedback of AES and teachers is more effective in correcting grammar and mechanical mistakes.

While efficiency is the promise of AI assessment (H. Zhang, 2021), not all student interviewees saw PE as an enormous added value. Opinions differed on the added value of grades, where some students mentioned it would help them better prepare for further assignments in the course and others mentioned a faster grade would not affect their style of work more. Similarly, time efficiency

and decrease of human bias were not seen as big enough problems that need drastic change or as individual benefits. Everything that would benefit students would positively affect their opinion on AI assessment, but not enough for students to advocate for the service. Seeing that students do not actually use the services this might explain why Performance Expectancy is not of great influence on this target group.

While students do not mention any of the themes as groundbreaking, it is important to note that the additional feedback and system transparency can improve student self-regulated learning skills and motivation to write (Selwyn, 2019; Ferguson, 2019).

The Motivator and Inhibitor factors were closely related. Especially the themes of trust and distrust showed an interesting dynamic. Trust is a key factor in technology acceptance and is enhanced through the user's faith in the interaction with the technology (Cyr, Hassanein, Head, & Ivanov, 2007; Hengstler, Enkel, & Duelli, 2016). Previous experiences with technology can affect an individual's level of trust, whether this is a lack of, bad, or good experience. In the context of students at the University of Twente, students clearly lack experience in AI assessment. It is crucial for acceptance to allow students to work with AI services that might improve their grade and not decide it. Building trust in AI assessment is the first step toward acceptance (Estevez, Garate, & Grana, 2019).

The possible replacement of teachers with AI services was also a point of discussion among the interviewees. Part of the student participants was adamantly against giving teacher tasks to AI services. This aligns with a study done by Bates, Cobo, Mariño, and Wheeler (2020), in which students experienced an emotional component to learning that cannot be satisfied by computer interaction only. Students in this study voiced that they want to be seen as individuals. In the current study, students mentioned that part of their reasoning to follow a course is because of the teacher. Garrison (2007) highlighted that this relational aspect of teaching can be achieved through technology with enough visual and communication support.

4.2 Teachers' Attitudes and Technology Preferences on AI Assessment

In general, teachers saw risks in the implementation of AI assessment. They did not have an outspoken positive opinion on the effect of AI assessment on the quality of education. Noteworthy, teachers that held a positive standpoint on AI technology thought that the implementation of AI assessment would not change the quality of education that much. In fact, most interviewees were convinced that their sense of judgment was better than that of a computer. This way of thinking heavily reduces willingness to adopt (Conijn, Kahr, & Snijders, 2021).

Overall, teachers expressed more enthusiasm to use AI assessment for formative assessment than for summative assessment. Teachers were less likely to give up their grading position to technology. Instead, they prefer to let AI services check students' writing in frequent assessments. Assigning certain tasks to AI assessment aligns with how automated assessment systems have been used in practice (H. Zhang, 2021; Molenaar, 2021). While AES has had more research in formative settings, it is still important that this type of automated assessment receives attention in summative contexts (West-Smith, Butler, & Mayfield, 2018).

One of the biggest concerns was that of transparency. Just like Kumar and Boulanger (2021) pointed out in their review, lack of transparency is a 'dealbreaker' for teachers in educational assessment. Teachers would not accept technology that would not explain how it would come to

a grade (i.e., black box approaches). For AI assessment to become acceptable, it is suggested that teachers need full insight into the inner rubric of the AI service, preferably through an easy-to-read and edit interface (Conijn et al., 2021).

Unlike students, teachers were more positive about the Performance Expectancy of AI assessment. This does not echo results from Jiang, Yu, and Wang (2020), where teachers' attitudes were negative toward the added value of AI assessment. Since teachers would be interacting closely with the services, they are looking for more beneficial aspects of the technology. Most importantly, the topic of time investment was a driver for this factor. Efficiency should be a selling point of AI assessment (Nazaretsky, Ariely, Cukurova, & Alexandron, 2022). Teachers are not keen on investing a lot of time into the implementation of AI services.

This does not align with the mentioned technology inexperience. Especially with AI technology, most teachers at the University of Twente have no experience and therefore, need help in getting used to the change. Not only is the lack of prior knowledge a challenge, but most problems also stem from the way that technology is designed. The apparent lack of involvement of educators in the design process of AI services is part of the problem (Bates et al., 2020). Instructional designers are advised to collaborate with teachers to ensure their demands are met. A human-centered design for AI programs facilitates interaction better for people with less technology experience (Riedl, 2019). Additionally, Vinichenko, Melnichuk, and Karácsony (2020) stated that for the sake of adoption, academic staff requires motivational methods that explain the benefits of AI assessment.

There were two different views on the role that AI services would play in assessment. In the interviews, it became clear that teachers would either view AI services as a supporting role or as a colleague. The difference between these two views is the level of trust teachers place in AI services. Teachers that saw AI assessment as a supporting tool were also more likely to also check their student assistants. Similarly, these teachers would not feel comfortable placing much responsibility on an AI tool. Even within this category, the level of support would differ per participant. This boils down to trust again, where the user is required to rely on the AI assessment. When the level of trust is not high enough, the capability of the AI service is distrusted (Abbass, 2019).

The teachers that did see AI services as a colleague were willing to accept its judgment, as long as it agreed with theirs. This did not mean that they would prefer summative AI assessments either.

This trust in AI assessment might also explain the purpose that teachers see for its use. Possibly due to the limited exposure to AI assessment services, most teachers found AI assessment best suitable for short text entry items. Just like students, teachers need more familiarity with AI technology before they are able to critically and practically consider AI assessment.

Inhibitors played a less crucial role in the teachers' acceptance of AI assessment. Perhaps because they do not experience the direct consequences of assessment and therefore experience less discomfort with technology. It was found in another study that Motivators had a stronger effect on technology usage than Inhibitors (Blut & Wang, 2020).

4.3 Student and Teacher Views

Both participant groups suffer from technology inexperience in the area of AI assessment. For most participants, the concept of AI assessment was alien and needed further explanation to be understood. This inexperience is a stepping stone toward distrust in new technology (Hengstler

et al., 2016; Nazaretsky et al., 2022). This explains why in both groups Perceived Risk was a predominantly present factor, which is in line with previous research (Teo & Liu, 2007; Warkentin et al., 2002).

As expected, Performance Expectancy was a positive factor in attitude toward technology, since both participant groups found efficiency to be an important consequence of AI assessment. Sohn and Kwon (2020) found that perceived value was a significant predictor of technology adoption in AI-based intelligent products. However, no matter the level of PE, in some cases interaction with a human is always preferred over technology (Kelly, Kaye, & Oviedo-Trespalacios, 2023). Similarly, both students and teachers are not fond of removing the human-in-the-loop in educational assessment.

However, student and teacher attitudes do differ on some points. It was found that students focus more on the direct consequences of AI assessment (e.g., time allocation, feedback, negative effects on their grades), whereas teachers focus more on the overall process (e.g., transparency, seeing AI as support or colleague). This might also explain why students had more present Inhibitor themes, as they would experience the consequences of AI assessment more severely.

Important to note is that students mostly saw the benefits of AI assessment if the teacher would invest this time back into the course by providing more feedback or spending more time in class. Teachers, on the other hand, agree that they already take enough time for all their courses and would put the extra time into other tasks.

4.4 Conclusion

The aim of this thesis was to uncover the relation of personal attitude and technology preference on the acceptance of automatic AI grading software in formal Higher Education. It was discovered that both factors of attitude and technology preference have an effect on the acceptance of AI assessment. Most importantly, in attitude it was found that Performance Expectancy and Perceived Risk are the strongest predictors of acceptance. Efficiency and outcomes expectations heavily influenced an individual's belief in AI assessment. For technology preference, trust was the most important predictor. Trust relied heavily on previous knowledge on technology and would be negatively influenced by a lack of knowledge.

Based on the results presented in this thesis, the following guidelines are suggested for instructional designers of AI assessment.

First, it is proposed that trust in AI assessment is increased. This can be done in two ways: increasing knowledge of AI assessment and exposure to positive experiences. This process is more of a gradual effect on the user population. Not all users can become AI experts, but the aim should be that people are able to judge the validity and possible errors of AI assessment results (Nazaretsky et al., 2022). Efficiency should be a selling point of AI assessment, but both teachers and students should be content with the effectiveness of AI assessment as well.

Second, it is recommended to implement AI assessment in formative settings first and then move toward summative assessment. Formative assessment brings fewer risks and thus fewer consequences for students. As teachers have mentioned, they like to remain in control over assessment and formative assessment tasks to increase efficiency (Zhu, Liu, & Lee, 2020). Some pointers for formative assessment are: interactivity with feedback prompts to increase learning gains (Correnti, Matsumura, Wang, Litman, & Zhang, 2022), content-specific feedback vs generic (Zhu et al., 2020) and transparency in grading and feedback decisions (West-Smith et al.,

2018). However, possibilities can be found over time to implement AI assessment in summative assessment as well. As Yıldız, İpek, and Gönen (2021) have pointed out, AI assessment can serve as a second rater while still providing the promise of efficiency and keeping the teacher in charge of grades.

Third, it is very important that full transparency is included in the system design. Both students and teachers have mentioned a need for transparency and explainability, albeit because of different wishes for both participant groups. Different users have different needs and to incorporate all necessary perspectives, it is important that teachers and students are involved in the design process (Clancey & Hoffman, 2021). The completeness of explanations is still debatable, as studies have found that there can be too much explainability in system decisions which causes over-reliance (Bussone, Stumpf, & O'Sullivan, 2015). Similarly, a human-in-the-loop approach is recommended, where teachers are given more control and room to intervene in AI assessment. This method can reduce technology-induced anxiety, prevent errors, and increase human agency (Nazaretsky et al., 2022; Galici, Käser, Fenu, & Marras, 2022).

4.5 Limitations

This study has been subject to limitations. First of all, the choice of theoretical framework might have been flawed. The UTAUT model has been used to measure the level of acceptance and willingness to adopt certain technologies. The student participant group in this study, however, will not directly use AI software for assessment, but merely be affected by it. While their willingness to accept the results of AI assessment is valid, it is hard to measure this with the UTAUT model since they do not affect the adoption process (Venkatesh et al., 2003). However, Conijn et al. (2021) stated that different stakeholders' views should be taken into account when determining design choices.

Secondly, while the combination of the UTAUT model and the TRI has been used in research before, there are some common ground areas that make it hard to exclusively code (e.g., Perceived Risk and Inhibitors) (Rahman et al., 2017; Cabrera-Sánchez et al., 2021). However, seeing that Perceived Risk is not originally part of the UTAUT model, this can be mitigated by returning to the original theoretical model. By heavily focusing on the current theoretical framework, external factors have not been taken into account (Y. L. Lai & Lee, 2020). It is therefore suggested that the current findings could be used for future research to broaden the perspective to other technology-related factors.

The third limitation is as Kelly et al. (2023) pointed out in their literature review: priori views on individuals' perception of technology acceptance before usage should not be confounded for their future behavior. How participants claim they would behave does not guarantee anything for future behavior. This is apparent It would be recommended to continue the study using existing AI assessment services to see if any conclusions from this thesis uphold.

The fourth limitation concerns the sample size of both participant groups. These were relatively small to conclude any significant findings within the technology baseline survey. The results of this study can therefore serve as an indication for future research. Aside from that, the sample consisted of participants from the same educational environment. Hence, the findings of this study might not be generalizable for other educational environments (Bornstein, Jager, & Putnick, 2013). For research into significance, it can be suggested to gather more participants and search for participants in other universities. However, seeing that this was an explorative study, the findings still uphold

their importance.

Similarly, while 20 participants completed the survey, only 12 were interviewed for further analysis. Hence, the qualitative analysis should be perceived with caution due to the low number of participants. Nonetheless, the interviews still provide valuable insights into AI assessment acceptance in student and teacher user groups.

4.6 Future Research

In sum, the current work presents a snapshot of AI assessment acceptance at the University of Twente. The logical next step is to further focus on broadening these insights through the actual usage of an AI assessment tool designed according to the provided guidelines. First, it is important to find out through which measures of trust in AI assessment can be improved and whether the suggested guidelines are effective (Conijn et al., 2021). Secondly, it is valuable to investigate the impact of AI assessment on the educational system. Thus, creating empirical evidence on the effectiveness and usefulness of AI assessment.

Accordingly, an exciting challenge would be to investigate factors that influence people on a social level. As Sohn and Kwon (2020) found in their research, subjective norm also has a significant effect on whether users choose to adopt AI technology. It would be interesting to see whether social norms also affect teachers and students in the area of AI assessment. For instance, how do the opinions of classmates change their opinion on the acceptance of AI assessment? These factors should be explored alongside a working prototype.

Instead of using the UTAUT model and Technology Readiness Index as predictive models, a quantitative method could be used to verify the findings of this study. Additionally, the generalizability of the findings would be increased through this method since larger samples can be used and findings can be compared with previous AI assessment-related research.

The effect of AI assessment on student performance and fairness is still a debatable factor of technology and should be subject to further research as well (Litman et al., 2021; G. Zhang et al., 2021). Further investigation into the actual consequences of AI assessment at the University of Twente could be a fine addition to the findings of this thesis.

The implementation process of an AI assessment tool has to be assessed. To see whether the University of Twente has a suitable environment for it to be implemented. The effectiveness of AI assessment is still the subject of research and it is important that the quality of education is still valued when AI assessment is implemented (Owoc et al., 2021).

Lastly, it is vital to understand the effects of AI assessment on the quality of education. The University of Twente is mostly a technical-oriented university, so this might affect the acceptance of AI assessment favorably. It would be vital to research whether this acceptance is reciprocated in other universities and educational systems of a different profile.

References

Abbass, H. A. (2019). Social Integration of Artificial Intelligence: Functions, Automation Allocation Logic and Human-Autonomy Trust. *Cognitive Computation*, *11*(2), 159–171. doi: 10.1007/s12559-018-9619-0

Adabi, A., & Berrada, M. (2018). Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence. *IEEE Access*, *6*, 52138–52160.

Aharony, N. (2015). Factors affecting the adoption of e-books by information professionals. *Journal of Librarianship and Information Science*, *47*(2), 131–144.

Amer, M., Daim, T. U., & Jetter, A. (2013). A review of scenario planning. *Futures*, *46*, 23–40. Retrieved from `http://dx.doi.org/10.1016/j.futures.2012.10.003` doi: 10.1016/j.futures.2012.10.003

Andrews, J. E., Ward, H., & Yoon, J. W. (2021). UTAUT as a Model for Understanding Intention to Adopt AI and Related Technologies among Librarians. *Journal of Academic Librarianship*, *47*(6), 102437. Retrieved from `https://doi.org/10.1016/j.acalib.2021.102437` doi: 10.1016/j.acalib.2021.102437

Arrieta, A. B., D\'\iaz-Rodr\'\iguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... others (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Bajaj, R., & Sharma, V. (2018). Smart Education with artificial intelligence based determination of learning styles. *Procedia Computer Science*, *132*, 834–842. Retrieved from `https://doi.org/10.1016/j.procs.2018.05.095` doi: 10.1016/j.procs.2018.05.095

Bates, T., Cobo, C., Mariño, O., & Wheeler, S. (2020). Can artificial intelligence transform higher education? *International Journal of Educational Technology in Higher Education*, *17*(1). doi: 10.1186/s41239-020-00218-x

Beseiso, M., Alzubi, O. A., & Rashaideh, H. (2021). A novel automated essay scoring approach for reliable higher educational assessments. *Journal of Computing in Higher Education*, *33*(3), 727–746. Retrieved from `https://doi.org/10.1007/s12528-021-09283-1` doi: 10.1007/s12528-021-09283-1

Blut, M., & Wang, C. (2020). Technology readiness: a meta-analysis of conceptualizations of the construct and its impact on technology usage. *Journal of the Academy of Marketing Science*, *48*(4), 649–669. doi: 10.1007/s11747-019-00680-8

Borenstein, J., & Howard, A. (2021, 2). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, *1*(1), 61–65. doi: 10.1007/s43681-020-00002-7

Bornstein, M. H., Jager, J., & Putnick, D. L. (2013). Sampling in developmental science: Situations, shortcomings, solutions, and standards. *Developmental review*, *33*(4), 357–370.

Bridgeman, B., Trapani, C., & Attali, Y. (2009). Considering fairness and validity in evaluating automated scoring. In *Annual meeting of the national council on measurement in education (san diego, ca, usa). ncme, mt. royal, nj* (pp. 1–18).

Brownell, E. (2022). Artificial Intelligence Impersonating a Human : The Impact of Design Facilitator Identity on Human Designers. (December). doi: 10.1115/1.4056499

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

Cabrera-Sánchez, J. P., Villarejo-Ramos, F., Liébana-Cabanillas, F., & Shaikh, A. A. (2021). Identifying relevant segments of AI applications adopters – Expanding the UTAUT2's variables. *Telematics and Informatics*, *58*(December 2019). doi: 10.1016/j.tele.2020.101529

Chang, A. (2012). UTAUT and UTAUT 2: A Review and Agenda for Future Research. *The Winners*, *13*(2), 106–114.

Chatterjee, S., & Bhattacharjee, K. K. (2020). Adoption of artificial intelligence in higher education: a quantitative analysis using structural equation modelling. *Education and Information Technologies*(January). doi: 10.1007/s10639-020-10159-7

Chatterjee, S., Rana, N. P., Khorana, S., Mikalef, P., & Sharma, A. (2021). Assessing Organizational Users' Intentions and Behavior to AI Integrated CRM Systems: a Meta-UTAUT Approach. *Information Systems Frontiers*(7491). doi: 10.1007/s10796-021-10181-1

Chen, H. (2022). Computer or Human : A Comparative Study of Automated Evaluation Scoring and instructors ' feedback on Chinese College Students ' English Writing. *Asian-Pacific Journal of Second and Foreign Language Education*, 1–21. Retrieved from https://doi.org/10.1186/s40862-022-00171-4 doi: 10.1186/s40862-022-00171-4

Clancey, W. J., & Hoffman, R. R. (2021). Methods and standards for research on explainable artificial intelligence: lessons from intelligent tutoring systems. *Applied AI Letters*, *2*(4), e53.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial intelligence*, *298*, 103503.

Conijn, R., Kahr, P., & Snijders, C. (2021). The Effects of Explanations in Automated Essay Scoring Systems on Student Trust and Motivation. , *xx*(2013), 1–18.

Cope, B., Kalantzis, M., & Searsmith, D. (2021). Artificial intelligence for education: Knowledge and its assessment in AI-enabled learning ecologies. *Educational Philosophy and Theory*, *53*(12), 1229–1245.

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, *3*(February), 100084. Retrieved from https://doi.org/10.1016/j.caeo.2022.100084 doi: 10.1016/j.caeo.2022.100084

Cugurullo, F. (2020). Urban Artificial Intelligence: From Automation to Autonomy in the Smart City. *Frontiers in Sustainable Cities*, *2*(July), 1–14. doi: 10.3389/frsc.2020.00038

Cyr, D., Hassanein, K., Head, M., & Ivanov, A. (2007). The role of social presence in establishing loyalty in e-service environments. *Interacting with computers*, *19*(1), 43–56.

Damerji, H., & Salimi, A. (2021). Mediating effect of use perceptions on technology readiness and adoption of artificial intelligence in accounting. *Accounting Education*, *30*(2), 107–130. Retrieved from https://doi.org/10.1080/09639284.2021.1872035 doi: 10.1080/09639284.2021.1872035

Doewes, A., Saxena, A., Pei, Y., & Pechenizkiy, M. (2022). Individual Fairness Evaluation for Automated Essay Scoring System.

du Boulay, B., Poulovasillis, A., Holmes, W., & Mavrikis, M. (2018). Artificial Intelligence And Big Data Technologies To Close The Achievement Gap.

Estevez, J., Garate, G., & Grana, M. (2019). Gentle Introduction to Artificial Intelligence for High-School Students Using Scratch. *IEEE Access*, *7*, 179027–179036. doi:

10.1109/ACCESS.2019.2956136

Ferguson, R. (2019). Ethical Challenges for Learning Analytics. *Journal of Learning Analytics*, *6*(3), 25–30.

Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. *Journal of Educational Measurement*, *59*(3), 288–313. doi: 10.1111/jedm.12333

Fridin, M., & Belokopytov, M. (2014). Acceptance of socially assistive humanoid robot by preschool and elementary school teachers. *Computers in Human Behavior*, *33*, 23–31.

Galici, R., Käser, T., Fenu, G., & Marras, M. (2022). Do Not Trust a Model Because It is Confident: Uncovering and Characterizing Unknown Unknowns to Student Success Predictors in Online-Based Learning. *LAK23: 13th International Learning Analytics and Knowledge Conference (LAK 2023), March 13â•fi17, 2023, Arlington, TX, USA*, *1*(1), 1–16. Retrieved from

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *Journal of Computer Assisted Learning*, *37*(5), 1207–1216. doi: 10.1111/jcal.12577

Garrison, D. R. (2007). Online community of inquiry review: Social, cognitive, and teaching presence issues. *Journal of Asynchronous Learning Networks*, *11*(1), 61–72.

González-Calatayud, V., Prendes-Espinosa, P., Roig-Vila, R., & Carpanzano, E. (2021). applied sciences Review Artificial Intelligence for Student Assessment: A Systematic Review. *Appl. Sci*, *2021*, 5467. Retrieved from `https://doi.org/10.3390/app` doi: 10.3390/app

GUO, F. (2009). Fairness of Automates Essay Scoring of GMAT AWA. *GMAC Research Reports*, *9*.

Gutierrez, S. S. M., Perez, S. L., & Munguia, M. G. (2022). Artificial Intelligence in e-Learning Plausible Scenarios in Latin America and New Graduation Competencies. *Revista Iberoamericana de Tecnologias del Aprendizaje*, *17*(1), 31–40. doi: 10.1109/RITA.2022.3149833

Han, T., & Sari, E. (2022). An investigation on the use of automated feedback in Turkish EFL students' writing classes. *Computer Assisted Language Learning*, 1–24.

Handoko, B. L., & Liusman, S. (2021). Analysis of external auditor intentions in adopting artificial intelligence as fraud detection with the unified theory of acceptance and use of technology (UTAUT) approach. *ACM International Conference Proceeding Series*, 96–103. doi: 10.1145/3481127.3481143

Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in education: Principles, policy & practice*, *4*(3), 365–379.

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices. *Technological Forecasting and Social Change*, *105*, 105–120.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... Koedinger, K. R. (2021). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*. doi: 10.1007/s40593-021-00239-1

Jaakkola, H., Henno, J., Lahti, A., Jarvinen, J. P., & Makela, J. (2020, 9). Artificial intelligence and education. In *2020 43rd international convention on information, communication and*

electronic technology, mipro 2020 - proceedings (pp. 548–555). Institute of Electrical and Electronics Engineers Inc. doi: 10.23919/MIPRO48935.2020.9245329

Jiang, L., Yu, S., & Wang, C. (2020). Second language writing instructors' feedback practice in response to automated writing evaluation: A sociocultural perspective. *System*, *93*, 102302. Retrieved from https://doi.org/10.1016/j.system.2020.102302 doi: 10.1016/j.system.2020.102302

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. *IJCAI International Joint Conference on Artificial Intelligence*, *2019-Augus*, 6300–6308. doi: 10.24963/ijcai.2019/879

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*(December 2022), 101925. Retrieved from https://doi.org/10.1016/j.tele.2022.101925 doi: 10.1016/j.tele.2022.101925

Kumar, V., & Boulanger, D. (2020). Explainable Automated Essay Scoring: Deep Learning Really Has Pedagogical Value. *Frontiers in Education*, *5*(October), 1–22. doi: 10.3389/feduc.2020.572367

Kumar, V., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, *31*(3), 538–584.

Kuo, K.-M., Liu, C.-F., & Ma, C.-C. (2013). An investigation of the effect of nurses' technology readiness on the acceptance of mobile electronic medical record systems. *BMC medical informatics and decision making*, *13*(1), 1–14.

Lai, M. L. (2008). Technology readiness, internet self-efficacy and computing experience of professional accounting students. *Campus-Wide Information Systems*, *25*(1), 18–29. doi: 10.1108/10650740810849061

Lai, Y. L., & Lee, J. (2020). Integration of Technology Readiness Index (TRI) Into the Technology Acceptance Model (TAM) for Explaining Behavior in Adoption of BIM. *Asian Education Studies*, *5*(2), 10. doi: 10.20849/aes.v5i2.816

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). *A Fairness Evaluation of Automated Methods for Scoring Text Evidence Usage in Writing* (Vol. 12748 LNAI). Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-030-78292-4$_2$1 doi: 10.1007/978-3-030-78292-4_21

Machicao, J. C. (2019). Higher education challenge characterization to implement automated essay scoring model for universities with a current traditional learning evaluation system. In *International conference on information technology & systems* (pp. 835–844).

Molenaar, I. (2021). Personalisation of learning: Towards hybrid human-AI learning technologies. *OECD digital education outlook*, 57–77.

Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers ' trust in AI-powered educational technology and a professional development program to improve it. (December 2021), 914–931. doi: 10.1111/bjet.13232

Ooge, J., Kato, S., & Verbert, K. (2022). Explaining Recommendations in E-Learning: Effects on Adolescents' Trust. In *27th international conference on intelligent user interfaces* (pp. 93–105).

Owoc, M. L., Sawicka, A., & Weichbroth, P. (2021). Artificial Intelligence

Technologies in Education: Benefits, Challenges and Strategies of Implementation. *IFIP Advances in Information and Communication Technology*, *599*, 37–58. doi: 10.1007/978-3-030-85001-2_4

Parameswaran, S., Kishore, R., & Li, P. (2015). Within-study measurement invariance of the UTAUT instrument: An assessment with user technology engagement variables. *Information and Management*, *52*(3), 317–336. doi: 10.1016/j.im.2014.12.007

Parasuraman, A. (2000). Technology Readiness Index (Tri): A Multiple-Item Scale to Measure Readiness to Embrace New Technologies. *Journal of Service Research*, *2*(4), 307–320. doi: 10.1177/109467050024001

Parasuraman, A., & Colby, C. L. (2015). An Updated and Streamlined Technology Readiness Index: TRI 2.0. *Journal of Service Research*, *18*(1), 59–74. doi: 10.1177/1094670514539730

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103–134.

Raffaghelli, J. E., Rodríguez, M. E., Guerrero-Roldán, A. E., & Bañeres, D. (2022). Applying the UTAUT model to explain the students' acceptance of an early warning system in Higher Education. *Computers and Education*, *182*(March 2021). doi: 10.1016/j.compedu.2022.104468

Rahman, S. A., Taghizadeh, S. K., Ramayah, T., & Alam, M. M. D. (2017). Technology acceptance among micro-entrepreneurs in marginalized social strata: The case of social innovation in Bangladesh. *Technological Forecasting and Social Change*, *118*, 236–245. Retrieved from `http://dx.doi.org/10.1016/j.techfore.2017.01.027` doi: 10.1016/j.techfore.2017.01.027

Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, *1*(1), 33–36. doi: 10.1002/hbe2.117

Rinjany, D. K. (2020). Does Technology Readiness and Acceptance Induce more Adoption of E-Government? Applying the UTAUT and TRI on an Indonesian Complaint-Based Application. *Policy & Governance Review*, *4*(1), 68. doi: 10.30589/pgr.v4i1.157

Roscher, R., Bohn, B., Duarte, M. F., & Garcke, J. (2020). Explainable Machine Learning for Scientific Insights and Discoveries. *IEEE Access*, *8*, 42200–42216. doi: 10.1109/ACCESS.2020.2976199

Selwyn, N. (2019). What's the problem with learning analytics? *Journal of Learning Analytics*, *6*(3), 11–19.

Sohn, K., & Kwon, O. (2020). Telematics and Informatics Technology acceptance theories and factors in fl uencing arti fi cial Intelligence-based intelligent products. *Telematics and Informatics journal*, *47*(December 2019), 1–14.

Somasundaran, S., Lee, C. M., Chodorow, M., & Wang, X. (2015). Automated scoring of picture-based story narration. *10th Workshop on Innovative Use of NLP for Building Educational Applications, BEA 2015 at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015*, 42–48. doi: 10.3115/v1/w15-0605

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., … Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*(May). doi: 10.1016/j.caeai.2022.100075

Taghipour, K. (2017). *Robust trait-specific essay scoring using neural networks and density estimators* (Unpublished doctoral dissertation). National University of Singapore (Singapore).

Teo, T. S., & Liu, J. (2007). Consumer trust in e-commerce in the United States, Singapore and China. *Omega*, *35*(1), 22–38. doi: 10.1016/j.omega.2005.02.001

Tsikriktsis, N. (2004). A technology readiness-based taxonomy of customers: A replication and extension. *Journal of service research*, *7*(1), 42–52.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*(2), 459–484. Retrieved from https://doi.org/10.1007/s41237-021-00142-y doi: 10.1007/s41237-021-00142-y

Uto, M., & Okano, M. (2021). Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases. *IEEE Transactions on Learning Technologies*, *14*(6), 763–776. doi: 10.1109/TLT.2022.3145352

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). Quarterly. , *27*(3), 425–478.

Vinichenko, M. V., Melnichuk, A. V., & Karácsony, P. (2020). Technologies of improving the university efficiency by using artificial intelligence: Motivational aspect. *Entrepreneurship and Sustainability Issues*, *7*(4), 2696–2714. doi: 10.9770/jesi.2020.7.4(9)

Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-Based System for Formative and Summative Assessment in Data Science Courses. *International Journal of Artificial Intelligence in Education*, *31*(2), 159–185. doi: 10.1007/s40593-020-00230-2

Warkentin, M., Gefen, D., Pavlou, P. A., & Rose, G. M. (2002). Encouraging Citizen Adoption of e-Government by Building Trust. *Electronic Markets*, *12*(3), 157–162. doi: 10.1080/101967802320245929

West-Smith, P., Butler, S., & Mayfield, E. (2018). Trustworthy automated essay scoring without explicit construct validity. *AAAI Spring Symposium - Technical Report*, *2018-March*, 95–102.

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding Blockchain Technology into IoT for Security: A Survey. *IEEE Internet of Things Journal*, *8*(13), 10452–10473. doi: 10.1109/JIOT.2021.3060508

Yıldız, H., İpek, S., & Gönen, K. (2021). The Use of an Automated Writing Evaluation System for Summative Assessment in an EFL Context: The Relationship Between Automated System Scores and Human Raters' Scores. *Dpublication.Com*, 143–153. Retrieved from https://www.dpublication.com/wp-content/uploads/2021/08/R36-4075.pdf

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019, 12). *Systematic review of research on artificial intelligence applications in higher education – where are the educators?* (Vol. 16) (No. 1). Springer Netherlands. doi: 10.1186/s41239-019-0171-0

Zhang, G., Raina, A., Cagan, J., & McComb, C. (2021). A cautionary tale about the impact of AI on human design teams. *Design Studies*, *72*, 100990.

Zhang, H. (2021). *Exploring Automated Essay Scoring Models for Multiple Corpora and Topical Component Extraction from Student Essays* (Unpublished doctoral dissertation). University of Pittsburgh.

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers and Education*, *143*(1). doi: 10.1016/j.compedu.2019.103668

A Informed Consent Form

The goal of this study This study aims to investigate the relationship between personal attitude toward technology and acceptance of artificial intelligence (AI) in Higher Education. In this webtool, you will be presented with a selection of cases in which AI will be used in an educational setting. Each scenario will focus on a different aspect of using AI in education, so your reaction might change depending on this factor. After completing the survey, there is a chance you might be invited for a follow-up interview to gain more in-depth information on your responses. This study has been reviewed and approved by the BMS Ethics Committee.

Method The following survey is divided into three parts:

- A demographic section
- A baseline technology acceptance questionnaire
- A selection of scenarios and your responses to those

The complete questionnaire takes approximately 20 - 30 minutes to complete. After completion of the questionnaire, you might be asked for an interview to gather in-depth information according to your given responses.

Confidentiality of data Due to the relevance of your responses for the follow-up interview, your personal information will not be anonymous to the researcher. The data collected will remain confidential and will be stored on a secured drive. Any personal data will not be used in the report and will be anonymized after analysis. The usage of this data will be strictly for academic purposes at the University of Twente. The retention period for the research data will be until the completion of the thesis. If you wish to access or erase any of your personal data, please contact the researcher (contact details below).

Voluntariness of participation Your participation in this study is entirely voluntary, so you can withdraw at any moment. This has no consequences. If you decide to withdraw from the study, your data will be used in the study up until the moment of withdrawal. If you wish to withdraw from the study or if you have any questions or remarks, feel free to contact the researcher (Sofie van den Berg) at: s.h.m.p.vandenberg@student.utwente.nl
Thank you for taking your time in assisting me with this study.

B Technology Acceptance Survey

B.1 Instruction

Technology acceptance level
Below you will be asked to answer a set of questions aiming to measure your technology acceptance level. Adjust the sliders (1-100) according to how well you agree or disagree with the statement (i.e., 1 would indicate totally disagree, 50 would indicate neutral, and 100 would indicate totally agree). This baseline consists of two models: Technology Readiness and the Unified Theory of Acceptance and Use of Technology.
Technology Readiness measures a person's general technology preferences, while UTAUT assesses attitudinal reactions to technology in a specific context.
Technology Readiness consists of two dimensions: Motivators and Inhibitors. UTAUT consists of three dimensions: Performance Expectancy, Perceived Risk, and Effort Expectancy.

B.2 Motivators

1.New technologies contribute to a better quality of education.

2.Technology makes me more efficient in my educational/professional life.

3.In general, I am among the first in my circle of friends to acquire new technology when it appears.

4.I keep up with the latest technological developments in the area of education.

B.3 Inhibitors

1.People are too dependent on technology to do things for them.

2.Whenever something gets automated, I need to carefully check that the system is not making mistakes.

3.If I use a high-tech product, I prefer to have a basic version over one with a lot of extra features.

4.I don't think technology should replace people in certain educational tasks (i.e., grading).

B.4 Performance Expectancy

1.Educational content assessed by AI technology is useful.

2.Educational content assessed by AI technology is effective.

3.Educational content assessed by AI technology is efficient.

B.5 Perceived Risk

1.Use of AI technology for assessing students is risky.

2.Use of AI technology for assessing students is confusing (i.e., non-transparent).

3.Use of AI technology for assessing students is not always correct (i.e., biased).

B.6 Effort Expectancy

1.I think using AI for grading would increase my productivity at work.

2.I would find it easy to learn how to use AI systems for grading.

3.I would find it easy to interact with AI systems while grading.

C Scenarios

C.1 Instruction

AI Scenarios
The next part includes a set of scenarios. You will be asked to carefully read through each scenario and answer a set of questions after. Be thorough when answering the questions and don't be afraid to put down any thoughts you have.

C.2 Student Scenarios

Language bias You are a student at the University of Twente. It is the last day of your deadline and you have worked hard to hand it in on time. At 12:00 you hear an email notification saying that your assignment has been graded. Last week, your professor excitedly announced that they would be using an AI program to grade the last assignments of the course, so all grades could be processed in time.
You eagerly click on the Canvas link. However, you notice that the grade doesn't reflect the amount of work you have put in. Instead, the email says you have, unfortunately, failed the assignment. A feedback document has been uploaded to your assignment as well. When you click on it, it presents you with your work and several auto-generated comments explaining why you have received or lost points. Upon further inspection, you notice that the AI software has deducted points on your use of language. Where your professor has always been quite lenient with using UK or US English, the AI software has solely been trained with UK English. Unfortunately, you have written your assignment in US English.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How would AI bias influence education if this service were to be implemented?
- What changes would this service need to be acceptable?

Limited transparency Normally, your professor grades your written assignments and provides you with feedback on how you can improve your grade. This type of feedback addresses both issues within the assignment as well as concepts that have been discussed in class. In the past, you have used this type of feedback to improve your work. However, you sometimes suspect that the professor's mood influences the way they grade (I.e., you know that when the professor is stressed, because of deadlines or evaluations, they tend to be stricter in their grading). With the introduction of the new AI service, all students receive their grades without any mood biases.
Yesterday, you handed in your assignment and your grade is available on Canvas. Because the feature is rather new, you choose to discuss the grades with your classmates. While the amount of work you have put into the assignment is reflected in your grade, there is no rubric available that explains how you've received and lost points. Your classmates also have not received any feedback on their grades.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How important is transparency for you in grading?
- What changes would this service need to be acceptable?

Strict guidelines Because your professor cannot provide everyone with an equal amount of feedback, the university has introduced AI software that can check your assignment and give you a predicted grade. The AI software makes use of a specific set of guidelines to check whether assignments are of a sufficient level. Much like the template your professor provides you with at the start of a course, the AI software bases your grade on the fulfillment of these guidelines. The use of this software is entirely optional for students.
You have worked on your assignment and are confident that you have put in all the necessary content for a good grade. However, you still want to try the AI software before handing it in. Upon uploading your assignment, you get a notification that the current state of your paper isn't satisfactory.
The problem: your paper does not follow the guidelines the AI has set and thus cannot be analyzed by the software. To get viable feedback and a predicted grade, your paper must exactly follow these guidelines and must be re-written.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How do you feel about the extra effort you must take to conform to the AI's standards?
- What changes would this service need to be acceptable?

General feedback Recently, AI software has been deemed as proficient to grade student exams – not only multiple-choice questions, but also open-type questions. Your course is a test run for this AI software. Nothing changes during the exam, except that you write it on a computer instead of on paper.
A few days after the exam, you receive your grade on Canvas. The grade is satisfactory and shows nothing strange. Accompanied with the grade is a link to your exam and the feedback the software has provided you with for your given answers – much like a viewing moment you would get if a professor had graded your exam.
You decide to look into the feedback with a couple of friends who have also taken the exam. Upon inspecting the feedback, you notice that the software gives good pointers on where you could have lost points. You and your friends find a handful of questions that you have all answered wrong – these answers differ between you all but come down to the same underlying error. The AI software has recognized this error and therefore provided you all with the same feedback prompt. As compared to feedback from your professor, this type of feedback is less personal and less applied to your specific answer.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- What do you look for when you receive feedback on an assignment? And does this service satisfy that need?
- What changes would this service need to be acceptable?

Transparency error Your professor has decided to let AI software go over your assignments for the rest of the course to increase the speed of providing students with grades. The software aims to provide the professor with suggestive grades accompanied with feedback that explains why points were granted or deducted. So, the AI software doesn't give a definitive grade. Upon handing in your assignment, you receive a document with the AI-generated feedback.
A day later you receive your grade, and you notice that your professor has not provided you with any additional feedback. It seems like your professor accepted all the AI-generated feedback and based your grade mostly on the predicted grade.
When you look closely at the feedback provided in the document, you notice that the software does not always draw the right conclusion. For example, sometimes it misunderstands your words and deducts points, while in other cases it rewards points.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How do you feel about AI software making decisions instead of simply informing them?
- What changes would this service need to be acceptable?

Double work This afternoon you had a lab session with a strict deadline. You are nervous, since this is your first lab report assignment, and you are quite unsure how to write one. At the end of the session, you hand in your lab report through Canvas, hoping for a good grade. In class, your professor announced that an AI agent would look through the lab reports and provide the professor with a grading decision.
The algorithm checks through every lab report and highlights areas that either need improvement or have been rewarded with full marks.
After two weeks, the grades are published on Canvas. You and your classmates wonder why it took so long and the professor explains that they had to go through all these grading decisions themselves to judge whether the AI software had made the right choice. All students receive a document with the AI-generated feedback.

- Would this scenario be acceptable for you?

- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How do you feel about the AI's efficiency in this scenario?
- What changes would this service need to be acceptable?

C.3 Teacher Scenarios

Rubric incompatibility The university has introduced AI software that can check student assignments and provide you with a predicted grade. The AI software makes use of a specific set of guidelines to check whether assignments are of a sufficient level. Much like the template you might give to your students at the start of a course, the AI software bases the grade on the fulfillment of these guidelines. The use of this software is entirely optional for students.
As you prepare the assignments for the course, you are requested to calibrate the AI software so it is compatible with your rubric. However, when you let the AI software run through your rubric and assignments, it gives you an error. Upon uploading the files, you get a notification that the current state of your files isn't compatible with the system.
The problem: the software works with internal settings that aren't compatible with your files. To get viable feedback and a predicted grade, you will have to re-write the rubric and assignment within the program, so it fits with the variables used in the AI software.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- In what case would you make use of an AI service like this?
- What do you think about the efficiency of a service like this?

Invisible rubric A new app has been integrated into Canvas at the start of the module – the university tell you it's new AI software that will help you grade student assignments. The software works just like the inbuilt plagiarism check on Canvas, so you don't have to install or use any other programs. For the software to work, the only requirement is to upload your rubric so that the AI knows how to grade the assignment.
In the third week of the course, your students hand in the first draft of their assignment. You decide this is the right time to test the AI software. Based on your rubric, the software goes through all the delivered assignments and provides you with a suggested grade per assignment. An hour after the deadline, you receive a notification that the software has analyzed all student assignments. As you open the provided log data, you notice that the software does not show how it has graded your students, but simply outputs a predicted grade.
Since the software is quite new, you decided to grade a random assignment all by yourself. When you compare the grades of that specific assignment, you notice that they are quite similar.

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How important is transparency for you with technology that, for example, helps you grade your students?
- In what case would you use a service like this?

Non-efficient interface The promise of AI technology is to increase overall efficiency - that's why the university has invested in AI software. For this academic year, the university has decided to finally implement the promised software and your course is one of the trial courses.
Just before the start of the module, you receive an email with an instructional manual and a short summary of what the software will do you for. In the long run, the AI will aid you in grading and assessing written assignments – which will save you a lot of time and you will still be in charge of the grading output.
However, as you scroll through the manual, you notice that the use of the software isn't as straightforward as you thought it would be. Instead, you think the User Interface might take some time to get used to (think: a couple of weeks in which you spend some hours figuring out the console).

- Would this scenario be acceptable for you?
- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How much interaction do you want with software like this? (i.e. being able to customize widgets, hands-off approach, etc.)
- In what case would you use a service like this?

Impersonal feedback In previous years, you went through all your students' assignments individually, providing them with personal feedback on a case-to-case basis. With the emergence of AI technology, the university has decided to increase the speed and efficiency of your grading work.
What's new? AI software has been designed and thoroughly tested to scan student assignments and provide suggestions for grades and feedback prompts. The feedback prompts all seem to be theoretically sound and correctly highlight specific parts of the students' assignments. You are free to accept or dismiss the prompts when using the program, as well as edit them.
If you choose to accept the feedback prompts, you notice that students will receive less personalized feedback which might hinder their personal development.

- Would this scenario be acceptable for you?

- If yes, what makes the scenario acceptable? If no, what aspect makes this scenario unacceptable?
- How will the quality of education be affected by the automation of tasks like these? And is it a negative change?
- In what case would you use a service like this?

D Interview Guide

Technology Readiness

- How compatible do you think AI is in the educational context/practice?
  - And how compatible is AI with educational values? (or your values)
  - What could make it more compatible with you or the educational context?

Motivators

- How much would you like to rely on AI in an educational setting?
- If an assessment AI of any sort runs for a trial period, would you try it out? (I.e., you have the option to get extra feedback on a paper before the deadline from the AI)
  - What would be your biggest turn-offs? Or pros?
  - What are your biggest hopes for AI?
- What would AI mean for your productivity/efficiency?

Inhibitors

- What are your main worries about implementing AI in education?
- Would you like to be dependent on technology like AI for grades? (and thus: passing a course, graduating)
- If AI developed itself in such a way that you could not easily understand the algorithm that is grading you, would that worry you?
- How do you think implementing AI will affect the quality of education?

Attitude

- To what extent do you think using AI for assessment in higher education is a good idea?

E Coding Scheme

E.1 Coding Scheme for Students

Factor	Theme	Definition	Frequency
Perceived Risk	Inaccuracy of assessment	The possibility that AI assessment could be inaccurate and affect student grades.	31
	No room for discussion	The inability to discuss assessment outcomes with the AI service.	23
	Inflexible grading guidelines	Grading guidelines that the AI service would uphold that would be inflexible or non-transparent.	20
	Decrease of the quality of education	The (negative) effect that AI assessment could have on the overall quality of education.	10
Performance Expectancy	Receiving feedback	The ability of AI assessment to provide students with feedback.	38
	Faster grading	The efficiency that AI services could bring when used in assessment, namely for the speed of grading.	12
	Decrease of human bias	The decrease of human-caused bias in educational assessment.	12
	Time efficiency of teachers	Due to AI assessment's efficiency, teacher would get more time for other educational related tasks.	10
Inhibitors	Lack of trust	Students show a lack of trust in the results of AI assessment.	22

Factor	Theme	Definition	Frequency
	Lack of human interaction	The lack of human interaction students would experience when AI services would take over (part of) assessment.	18
	Dependence	Students become dependent on AI services for their grades and feedback.	11
Motivators	Trust	Students show levels of trust in the results of AI assessment.	12
	Technology potential	Students see potential in AI services to enhance assessment.	7

E.2 Coding Scheme for Teachers

Factor	Theme	Definition	Frequency
Perceived Risk	Quality of education	The effects that AI assessment has on the quality of education.	15
	Risk of bias	Risks of bias that AI assessment could bring and annihilate in education.	17
	Transparency issues	Issues with (lack of) transparency that AI services cause when used for assessment.	12
	Replacement of staff	The effect of staff replacement that AI assessment brings for teachers.	4
Performance Expectancy	Positive time management	The time that is freed up for teachers when AI assessment takes over some tasks.	23
	AI in a supporting role	How teachers see AI assessment as a supporting factor in their work.	20
	AI for specific assessment types	Different types of assignments that teachers see fit for AI assessment.	17
Effort Expectancy	Inexperienced technology usage	The inexperience of teachers with technology usage, linked to AI services.	9
Inhibitors			9
Motivators	Trust	Teachers show levels of trust in the results and usage of AI assessment.	14
	Seeing AI as a colleague	Teachers see AI services as colleagues when used for assessment.	5